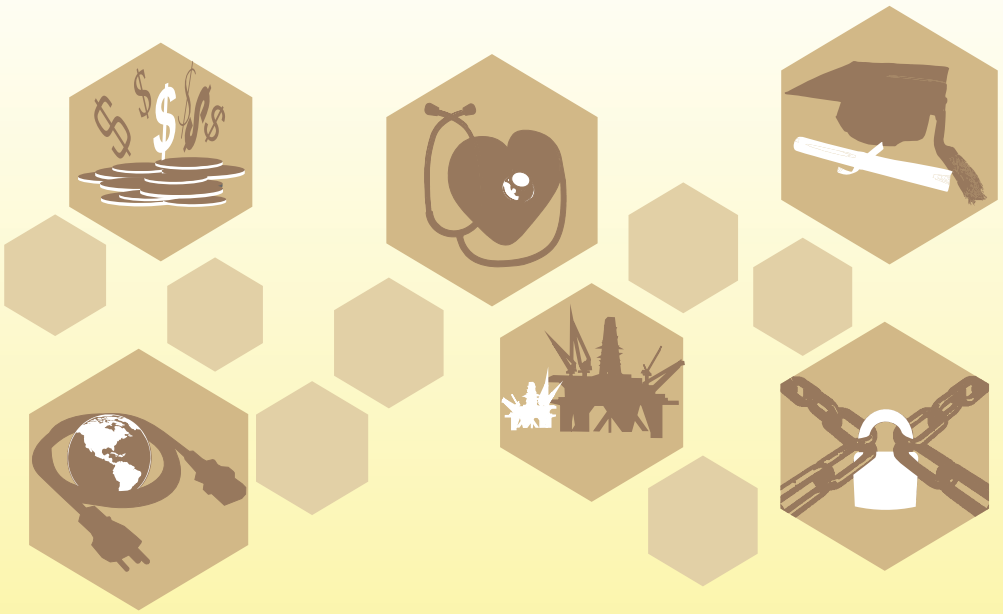
GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL

Relatos do 3º Seminário

Organizadoras
Ana Carolina Salgado, Claudia Lage Rebello da Motta, Flávia Maria Santoro



SBC 35 ANOS
Sociedade Brasileira de Computação

Grandes Desafios da Computação no Brasil

Relatos do 3º seminário

Organizadoras:
Ana Carolina Salgado
Claudia Lage Rebello da Motta
Flavia Maria Santoro

Promoção:
Sociedade Brasileira de Computação

Apoio:
Ministério da Ciência, Tecnologia e Inovação (MCTI)

Associoção Brasileira das Empresas de Tecnologia da Informação e Comunicação **(**Brasscom)

Patrocínio:
EMC

Realização:
Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais
Universidade Federal do Rio de Janeiro

Sociedade Brasileira de Computação

CONSELHO SUPLENTE
Mandato 2013-2015
Daltro José Nunes (UFRGS)
Rodolfo Jardim de Azevedo (UNICAMP-IC)
Aline Maria Santos Andrade (UFBA)
Karin Koogan Breitman (PUC-Rio)
Alessandro Fabrício Garcia (PUC-Rio)

COMISSÃO DE EDUCAÇÃO
Diretora: Mirella Moura Moro (UFMG)
Membros:
André Costa Drummond (UnB)
Carina Friedrich Dorneles (UFSC)
Ecivaldo de Souza Matos (IFSP)
Jair Cavalcante Leite (UFRN)
Renata Mendes de Araujo (UNIRIO)
Ronaldo Celso Messias Correia (UNESP)
Simone de Lima Martins (UFF)
Tayana Uchoa Conte (UFAM)

EQUIPE ADMINISTRATIVA
Supervisão Administrativa: Adriana Leandro Nowicki
Supervisão Financeira e Contábil: Fernanda dos Santos Jorge
Comunicação: Sílvia Carolina Costa Neves de Matos, Franciele Kettermann
Eventos: Pâmela Cilene Azevedo de Oliveira
Contas a Pagar: Fernanda Coimbra Maggioni
Faturamento: Juliana Pinto Betat
Secretária: Vanessa Vargas
Operacional: Emanuelle Quadros dos Santos
Suporte Técnico: Felipe da Silva Formiga

3º Seminário Grandes Desafios da Computação no Brasil – Fase 2

COMITÊ DE PROGRAMA
Altigran Soares da Silva (UFAM)
Ana Carolina Salgado (UFPE)
Augusto Sampaio (UFPE)
Avelino Zorzo (PUCRS)

Claudia Cappelli (UNIRIO)
Claudia Lage Rebello da Motta (UFRJ)
Cleidson de Souza (UFPA)
Daniela Barreiro Claro (UFBA)
Flavia Maria Santoro (UNIRIO)
Flavio Wagner (UFRGS)
Jonice Oliveira (DCC/IM/UFRJ)
José Maldonado (SSC/ICMC-USP/São Carlos)
Marcos Borges (UFRJ)
Rafael Prikladnicki (PUCRS)
Renata Araujo (UNIRIO)
Renato Cerqueira (IBM Research - Brazil)
Thais Vasconcelos Batista (UFRN)

COMITÊ ORGANIZADOR
Claudia Lage Rebello da Motta (Coordenação Geral)
Flavia Maria Santoro (Coordenação Geral)
Ana Lucia Rodrigues (Coordenação Executiva)
Adriano Barros
Alisson Maldaner
Avelino Zorzo
Jade Soares
Juarez Castro
Mônica Machado Ribeiro
Roselinda Passos
Ricardo Lacerda Caiado
Tania Cristina Oliveira

REDAÇÃO DO RELATÓRIO
Clara Rescala

A série de Seminários Grandes Desafios da Computação no Brasil, organizada pela Sociedade Brasileira de Computação (SBC), tem sido uma iniciativa pioneira no sentido de planejar e direcionar a pesquisa em Computação para um período de 10 anos (2006 a 2016). O impacto positivo abriu portas para o lançamento de editais de fomento à pesquisa e organização de eventos em torno dos temas, além de influenciar ações mais abrangentes como definir os Grandes Desafios de Pesquisa em Computação para América Latina.

Na sequência da definição de temas, o foco passou a ser os domínios de aplicação e a integração com indústria e governo. Os grandes desafios foram revisitados na perspectiva de agricultura, transporte, educação, indústria de Tecnologia da Informação e Comunicação, energia, aeronáutica, defesa, meio ambiente, bioenergia, biodiversidade, cidadania, governo eletrônico, saúde, entre outros, resultando em uma caracterização matricial: em uma dimensão, os grandes problemas de pesquisa do ponto de vista da Ciência da Computação e na outra, as aplicações desafiadoras e estratégicas para o País.

Mais além, a SBC trouxe como motivação para o 3º Seminário a ideia de promover redes de colaboração temáticas em função de problemas reais envolvendo os diferentes segmentos: governo, indústria e academia. Para isso, foram identificadas através de chamada de trabalhos parcerias possíveis dentro do contexto dos grandes desafios da Computação destacados nos anos anteriores.

Este livro relata e registra o resultado de todo este esforço, conquistado através do trabalho de vários grupos envolvidos. Entre artigos científicos, propostas de projetos e discussões em workshops, palestras e painéis, muito conteúdo é disponibilizado para servir como base para novas propostas para os próximos 10 anos que virão.

Agradecemos à Diretoria da SBC pela oportunidade e confiança na organização e realização deste evento. Agradecemos também a Brasscom - Associação Brasileira das Empresas de Tecnologia da Informação, e Comunicação e ao Ministério da Ciência, Tecnologia e Inovação, pelo apoio à realização da Fase 1 e da Fase 2, respectivamente. As experiências e conhecimentos adquiridos são compartilhados com a comunidade aqui!

A todos, uma boa leitura.

Ana Carolina
Claudia
Flavia

11.1 ARTIGOS

EDUCAÇÃO

SISTEMA BANCÁRIO/FINANCEIRO

PETRÓLEO/ENERGIA

1. HISTÓRICO DOS SEMINÁRIOS REALIZADOS PELA SBC

1.1 PRIMEIRO SEMINÁRIO GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL

O Primeiro Seminário Grandes Desafios da Computação no Brasil[1], organizado pela SBC, foi realizado em São Paulo nos dias 8 e 9 de maio de 2006. Esta foi uma iniciativa pioneira em Computação no país, no sentido de planejar e direcionar a pesquisa em Computação para um período de 10 anos (de 2006 a 2016). O seminário reuniu durante os dois dias 26 pesquisadores brasileiros da área de Computação.

O impacto positivo desta iniciativa tem sido bastante significativo, pois permitiu identificar grandes temas de pesquisa para o período de uma década, lançar editais de fomento à pesquisa direcionados para os temas identificados, organizar eventos em torno dos temas e, inclusive, influenciar ações mais abrangentes como a definição dos Grandes Desafios de Pesquisa em Computação para América Latina.

Cinco desafios foram identificados no primeiro seminário:

1. Gestão da Informação em grandes volumes de dados multimídia distribuídos
2. Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem natureza
3. Impactos para a área da Computação da transição do silício para novas tecnologias
4. Acesso participativo e universal do cidadão brasileiro ao conhecimento
5. Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos

1.2 SEMINÁRIO GRANDES DESAFIOS DA COMPUTAÇÃO PARA AMÉRICA LATINA - CHARLA

O seminário Grandes Desafios da Computação para América Latina, CHARLA, foi realizado em Buenos Aires nos dias 5 e 6 de setembro de 2008. Este evento teve como objetivo a discussão de desafios com a comunidade latino-americana, e teve como resultado a identificação de quatro grandes desafios:

1. Tecnologias de Informação e Comunicação Orientadas ao Cidadão
2. Multilinguismo e Identidade Latinoamericana em um Mundo Digital
3. Computação orientada ao monitoramento e controle ambiental
4. Redes Colaborativas Complexas (na América Latina)

[1]http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=view.download&catid=50&cid=11

1.3 SEGUNDO SEMINÁRIO GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL

O objetivo geral do Segundo Seminário Grandes Desafios da Computação no Brasil[2], realizado em Manaus nos dias 3 e 4 de março de 2009, foi fortalecer a pesquisa em torno dos desafios para a próxima década, focando na integração com a indústria de Tecnologia da Informação e Comunicação (TIC), detalhando desafios existentes ou propondo novos desafios. Os resultados foram sintetizados em torno dos seguintes temas:

1. Redes Complexas de Colaboração e Gestão da Informação sobre Grandes Volumes de Dados

2. Modelagem Computacional de Sistemas Complexos Artificiais, Biológicos e Inspirados na Natureza

3. Impactos para a computação devido à evolução e heterogeneidade tecnológicas de implementação do hardware

4. Grandes Desafios em Computação Aplicada e Entendendo a Web Desenvolvimento de sistemas confiáveis.

O ponto central da reunião de Manaus foi a revisão dos grandes desafios na perspectiva de domínios de aplicação, resultando em uma caracterização matricial dos grandes desafios:

- Agricultura
- Transporte
- Educação
- Indústria de TIC
- Energia
- Aeronáutica
- Defesa
- Meio ambiente
- Bioenergia
- Biodiversidade
- Cidadania
- e-gov
- Saúde

[2]http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=view.download&catid=50&cid=237

Em uma dimensão, o foco centrou-se nos grandes problemas de pesquisa do ponto de vista da Ciência da Computação e na outra, a ênfase centrou-se em aplicações desafiadoras e estratégicas para o País e os problemas que essas trazem para cada um dos Grandes Desafios.

2. TERCEIRO SEMINÁRIO GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL – FASE 1

O terceiro seminário teve como objetivo identificar os grandes desafios reais na área de Tecnologia da Informação e Comunicação do setor e de grandes instituições públicas brasileiras. A ideia central desta Fase 1 foi selecionar algumas das áreas estratégicas indicadas nos Ecossistemas Digitais do TI Maior – Programa Estratégico de Software e Serviços de Tecnologia da Informação, proposto pelo MCTI, e promover a discussão conjunta indústria e academia. Também foram convidados os dirigentes de quatro centros de P&D de empresas estrangeiras que estão instalados no Brasil.

O objetivo principal, em suma, foi possibilitar a apresentação dos desafios da área de TI pelas instituições representadas no evento e promover redes de colaboração temáticas em função de problemas reais que envolvam os diferentes segmentos: governo, indústria e academia. Como resultado final espera-se a elaboração de projetos conjuntos entre estes segmentos.

Estavam presentes no seminário 11 empresas e 12 universidades (20 pesquisadores), totalizando cerca de 50 pessoas. Algumas informações sobre o evento estão disponíveis no site da SBC, nas notícias em destaque[3] e no Boletim de Notícias[4]. Também no site da Brasscom encontramos uma descrição geral do evento[5], as apresentações dos palestrantes de cada painel[6] e alguns relatos sobre o evento[7].

Este relatório é apresentado por meio de relatos específicos por painel do evento, seguido por considerações sobre os principais temas da computação identificados nas discussões de todo o evento.

[3]http://www.sbc.org.br/index.php?option=com_content&view=article&id=1126:3o-seminario-grandes-desafios-da-computacao-no-brasil-integrou-setores-para-debaterem-as-principais-demandas-das-areas-de-inovacao-e-tecnologia-da-informacao-e-comunicacao-&catid=65:destaques
[4]http://www.sbc.org.br/index.php?option=com_content&view=article&id=1143&Itemid=945
[5]http://www.brasscom.org.br/brasscom/Portugues/detNoticia.php?codArea=2&codCategoria=25&codNoticia=424
[6]http://www.brasscom.org.br/brasscom/Portugues/detInstitucional.php?codArea=3&codCategoria=22
[7]http://www.brasscom.org.br/brasscom/Portugues/detNoticia.php?codArea=2&codCategoria=25&codNoticia=42

2.1. PAINEL 1 – SISTEMA BANCÁRIO/FINANCEIRO

Este painel foi composto pelos seguintes palestrantes: Haroldo Jayme Martins Froes Cruz - Chefe Adjunto do Departamento de TI do Banco Central do Brasil; Anderson Luis Cambraia Itaborahy – FEBRABAN / Banco do Brasil; Claudio Vita Filho – Vice -presidente de Novos Negócios da ITAUTEC; e Gustavo Roxo – sócio da Consultoria Booz&Co.

Todos os palestrantes estavam de acordo com aspectos como conectividade, mobilidade e segurança no grande volume de dados e de transações gerados pelas aplicações bancárias. Foi falado sobre a elevação da maturidade dos processos de governança de TI nos bancos. As aplicações bancárias carecem de soluções mais inteligentes e complexas e o desenvolvimento dos sistemas deve ter ciclos de vida mais curtos. A segurança e a qualidade da informação são dos principais aspectos a serem considerados neste tipo de aplicação, tanto para redução de fraudes como para a garantia na legitimidade das comunicações com o uso de mensagens criptografadas e com assinaturas digitais. No entanto, a segurança não pode prejudicar a usabilidade, o cliente está cada vez mais exigente com relação à eficiência operacional das aplicações funcionando em todos os canais de acesso: no banco, no site e em seus próprios equipamentos móveis.

O grande volume de dados gera a necessidade de análises das transações assim como análises de risco de crédito. Tecnologias associadas a Big Data serão fundamentais para as análises bancárias.

Um outro ponto discutido foi com relação ao uso da computação em nuvem nos bancos. Existem questões legais que independem de TI, e uma solução seria o uso de nuvens privadas. Além disso, muitos investimentos foram feitos pelos bancos em datacenters.

Por fim, foi discutido que ainda se tem muito 'papel' circulando nos bancos e que além da maturidade dos processos de governança, a TI tem um papel fundamental para transformar e alavancar a eficiência bancária.

Comentários Gerais do Painel:

- Computação em Nuvem: por que não são usadas nos bancos? Questões de direito que independem de TI, solução seria o uso de nuvens privadas. Além disso, muitos investimentos foram feitos pelos bancos em datacenters.
- Cliente cada vez mais exigente
- Por que ainda se tem movimentação de papel nas agências? Está faltando Software!!
- A Gestão da TI dentro dos bancos precisará operar como um negócio dentro do negócio.

2.2. PAINEL 2 – ÁREAS CHAVE [PETRÓLEO, ENERGIA E DEFESA]

Neste painel participaram como palestrantes: Marcelo Gattass – Pesquisador responsável pelo TecGraf – PUC-Rio, substituindo o palestrante da Petrobrás; Carlos Vinicius Frees – Especialista em projetos de TIC da ABDI; Sergio Aguiar – Gerente de Arquitetura Empresarial da Embraer; e o Gen. Div. Jose Carlos dos Santos – Comandante do Centro de Defesa Cibernética.

Os temas discutidos neste painel e seus principais desafios por área são apresentados a seguir.

PETRÓLEO E GÁS

Sobre esta área foram discutidos alguns desafios e, em particular, os da Petrobrás. Para o desenvolvimento das aplicações neste domínio são necessários fundamentalmente conhecimentos de geoprocessamento, computação gráfica, realidade virtual, Web, entre outros. Exemplos de aplicações próprias da Petrobrás: visualização e interpretação de dados sísmicos, modelagem geológica de sistemas petrolíferos, visualização e gerenciamento de reservatórios, modelagem em engenharia e geologia, geomecânica computacional, realidade aumentada e interatividade digital, engenharia de sistemas distribuídos, jogos de treinamento e simulação.

Big Data foi um tema citado como relevante para predição em toda a cadeia de óleo & gás e, em especial, exploração, produção e logística. Não se pode deixar de falar em qualidade dos dados e informações e em problemas de integração de dados heterogêneos.

ENERGIA

Um dos principais pontos discutidos foi a inteligência para eficiência energética (Smart Grids), ou seja, viabilizar a implantação de redes elétricas inteligentes e tecnologias associadas; incentivar a exportação de soluções; incentivar novos modelos de negócio; ampliar e desenvolver a cadeia produtiva nacional de semicondutores, displays, equipamentos eletroeletrônicos, e softwares e serviços; e modernizar a infraestrutura de telecomunicações associada à rede elétrica. Para isso, serão necessários componentes digitais estratégicos (smart meters). Para garantir estes serviços serão necessários novos paradigmas para computadores seguros e resilientes, métodos eficazes para garantir segurança de código e sistemas de alta confiabilidade.

Um tema fundamental na discussão é o de Cidades inteligentes que envolve: serviços públicos, serviços de utilidade pública, mobilidade urbana, prevenção de desastres, segurança, educação, saúde, comunicação e informação. Estes serviços dizem respeito a: transporte, água, energia, telecomunicações, saneamento, gás e edificações. Foram citados como elementos centrais: o cidadão, os negócios e a governança pública. Quanto à governança pública os principais aspectos de TI citados foram: Sistemas de Gestão, Sistemas de Controle e Monitoramento, Big Data e Computação em nuvem, Computação Ubíqua e Processamento Intensivo, Sistemas analíticos e de predição, Sensoriamento e dispositivos inteligentes.

DEFESA

Muitos desafios foram apontados pelo representante da Embraer[8]. Os mesmos são listados abaixo associados a áreas específicas da computação:

Fusão de dados de alto nível/ Análise de intenções: Métodos computacionais para análise de intenções de agentes em ambientes complexos, constituídos por diversas fontes de informação.

Defesa/Resiliência Cibernética: Novos paradigmas para Computadores Seguros, Resilientes e Adaptativos; Soluções de segurança cibernética em redes virtuais na nuvem; Super computação; Computação de alto desempenho; Métodos eficazes para garantir segurança de código; Métodos eficientes para criptografia completamente homomórfica (Full Homomorphic Encryption)

Sistemas de Alta Confiabilidade: Métodos eficazes e eficientes para garantir alta confiabilidade de código, no sentido de funcionamento correto e segurança para o usuário e do sistema.

Inteligência Artificial/Aprendizagem de Máquinas: Desenvolver métodos para tornar operacional o novo paradigma de Programação Genética para programação de sistemas de aprendizagem de máquinas.

Sistemas Distribuídos / Estimação Distribuída: Desenvolver formas eficientes do ponto de vista computacional e de custo de comunicação para estimar de forma distribuída o estado de objetos de interesse com base em uma rede de sensores/ agentes.

[8]http://www.brasscom.org.br/brasscom/Portugues/detInstitucional.php?codArea=3&codCategoria=22

Processamento de Imagens: Reconhecimento de padrões em imagens.

Sistemas de Sistemas: Métodos e algoritmos inovadores para otimização operacional através do gerenciamento da saúde de sistemas, e do conhecimento do estado atual e predição do estado futuro da condição dos equipamentos.

2.3. PAINEL 3 – SAÚDE

Os seguintes palestrantes discutiram a área de saúde: Augusto Cesar Gadelha – Diretor do DATASUS / Ministério da Saúde; Dr. Flavius Augusto Albieri – Assessor Técnico de Gabinete da Secretaria Municipal de Saúde de SP; Dr. Marco Antonio Gutierrez – Presidente da Sociedade Brasileira de Informática em Saúde; e Dra. Marcia Ito – Médica pesquisadora do IBM Research.

Um dos principais pontos discutidos foi o Prontuário Eletrônico do Paciente (PEP). O PEP tem que ser eficaz para facilitar a consulta médica com mais usabilidade, mais integridade e mais privacidade. Para o PEP é necessária a integração dos sistemas com uma visão unificada de dados do paciente, atendendo aos padrões nacionais (TISS) e internacionais (HL7, IHE, DICOM, SNOMED-CT, CID10...) para interoperabilidade semântica entre sistemas. O acesso ao PEP, inclusive em equipamentos móveis, deve ser controlado a partir de perfis dos usuários.

Para o desenvolvimento das aplicações médicas são necessárias redes de conectividade, aplicações de telemedicina, portais de saúde do cidadão, comunicação móvel, uso de SMS para gerenciar no-show, RFID, infra-estrutura de chaves públicas. São aplicações de alta disponibilidade com transmissão de sinais e parâmetros vitais em tempo real, transmissão, armazenamento e visualização de imagens médicas e conexão com equipamentos médicos e de beira-de-leito, levando ao desenvolvimento de sistemas embarcados.

O grande volume de dados aponta para a necessidade de aplicações de Business Intelligence (BI): sistemas especialistas, redes neurais artificiais e data mining. Além disso, as tecnologias associadas a Big Data para análise de dados em larga escala e para oferecer indicadores clínicos e assistenciais. Também nestas aplicações não se pode deixar de falar em segurança da informação e de ética.

Um dos pontos relevantes apontados foi o aumento da expectativa de vida da população que leva à elevação das doenças crônicas no BR. Em 2015 os sistemas de saúde de vários países não serão sustentáveis, pois os doentes crônicos demoram a morrer... Será necessária então a gestão de crônicos, ou seja, sistemas de informação para cuidar do paciente crônico de forma centrada nele, para assistir e monitorar o paciente ao longo do tempo.

O representante do Hospital do Coração[9] resume o que considera como principais desafios de TI na área de saúde:

• Soluções inovadoras para gargalos computacionais (redes, armazenamento e processamento);

• Data warehouses/ Data Mining em tempo-real (medicina personalizada, apoio à decisão, alertas);

• Dispositivos móveis e vestíveis mudando a coleta e visualização de informações no ponto-de-cuidado, e redefinindo processos;

• Processamento de linguagem natural e gestão de ontologias (interoperabilidade semântica);

• Sistemas integrados no contínuo da saúde (homecare, atenção primária ambulatorial, hospitais e clínicas)

Comentários Gerais do Painel:

- Regulação de padrões de laudos e atendimento gratuito ao cidadão.
- O grande desafio é dar saúde de qualidade ao cidadão gratuitamente.
- Poucas pessoas trabalhando com pesquisa em TIC e medicina

2.4. PAINEL 4 – NOVOS MODELOS CONTRATUAIS DE P&D

Neste painel participaram como palestrantes: Paulo Mól – Diretor de Inovação da CNI; André Castro Pereira Nunes – Chefe do Depto. de Tecnologia da Informação e Serviços da FINEP; Ricardo Rivera de Sousa Lima – Gerente do Deptº de TI e Comunicação da Área Industrial do BNDES; e Ricardo Gonzaga – Representante da Agência Brasileira de Desenvolvimento Industrial.

Comentaremos a seguir o posicionamento de cada um dos palestrantes, finalizando com comentários gerais da plateia.

CNI

Quanto à formação de recursos humanos foi falado da importância de se investir no ensino técnico profissionalizante e no ensino superior com o incentivo à formação de engenheiros. Também foi falado sobre a importância de se trabalhar mais a PI nas empresas. Outro ponto citado foi quanto a estimular a internalização de empresas, assim como trazer empresas internacionais para o Brasil. Deve ser aumentado

[9] http://www.brasscom.org.br/brasscom/Portugues/detInstitucional.php?codArea=3&codCategoria=22

o incentivo à P&D e à inovação na empresa privada, convencendo os empresários a gastar mais em P&D.

Um outro ponto relevante foi a fala sobre a Embrapii (Empresa Brasileira de Pesquisa e Inovação Industrial). A criação da Embrapii foi motivada pela lacuna de financiamento existente entre a pesquisa básica (laboratórios), e a comercialização, chamado "vale da morte da inovação". Como financiar o precompetitivo de forma mais ágil com foco em inovação da demanda empresarial? Hoje a Embrapii conta com a participação de três institutos: Senai/Cimatec – Centro Integrado de Manufatura e Tecnologia, IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo, e INT – Instituto Nacional de Tecnologia. A Embrapii não é uma empresa, não tem prédio, é uma estrutura de financiamento enxuta e ágil para repasse de recursos para os institutos, e acompanhamento dos projetos.

FINEP

A definição de inovação dada pela FINEP:

"Inovação é a introdução de novidade ou aperfeiçoamento no ambiente produtivo ou social que resulte em novos produtos, processos ou serviços. Inovações devem, necessariamente, estar disponibilizadas no mercado, aplicadas nas organizações ou transferidas para a sociedade. A inovação pode apresentar escala local, nacional ou mundial. Pode ser incremental ou radical".

Por que inovar?

- Diferencial competitivo para as empresas na atual economia globalizada.
- Importante instrumento de desenvolvimento e transformação econômico-social.

Objetivo da FINEP

O objetivo da FINEP é atuar em toda a cadeia da inovação, com foco em ações estratégicas, estruturantes e de impacto para o desenvolvimento sustentável do Brasil.

Áreas prioritárias: Defesa, aeroespacial e aeronáutica, Tecnologia da Informação e Comunicação, Energias renováveis, Tecnologias limpas, Petróleo e Gás, Novos materiais, Saúde, Desenvolvimento social e tecnologias assistivas, Biotecnologia, Nanotecnologia.

Oportunidades nas Áreas de TICs:

Interatividade: TV digital, jogos digitais, realidade Aumentada, 3D e tecnologias 3D 4k, convergência de mídias

Infraestrutura de Telecomunicações: apoio ao desenvolvimento de equipamentos de tecnologias sem fio de alto rendimento (LTE - Long Term Evolution), internet das coisas (Ipv6 - Internet Protocol version 6).

Plataforma de Serviços: aplicativos voltados para os eventos esportivos, soluções visando a segurança da informação e criptografia

Microeletrônica: design, fabricação de circuito integrado, encapsulamento e teste

Computação em Nuvem: plataforma - PaaS –aplicativos para provedores de serviços na nuvem (virtualização, segurança, gestão); infraestrutura - IaaS – datacenter para serviços na nuvem.

BNDES

"Inovação em TICS está na linha principal de financiamento do BNDES".

O BNDES acompanha o esforço do País para estimular a Inovação. Foram apresentados os principais programas, linhas e produtos do banco nos últimos anos para a inovação. Especialmente foram citados os programas Prosoft, precursor da inovação no BNDES, Criatec, fundo de capital semente, e Funtec, apoio não reembolsável para ICTs. Além disso, existem planos conjuntos de apoio à inovação envolvendo além do BNDES, a FINEP e outros agentes financiadores.

Foi dado destaque para o Inova Empresa para 2013-2014. O Inova Empresa, que tem TIC como uma das áreas prioritárias, tem como objetivo o investimento em inovação para elevar a produtividade e a competitividade da economia brasileira: ampliando o patamar de investimentos e do apoio a projetos de risco tecnológico e fortalecendo as relações entre empresas, ICTs e setor público.

ABDI

A ABDI é responsável pela política industrial no Brasil. Dividida em setores entre eles TIC, área médica, e, em especial, os setores prioritários do plano Brasil Maior. Participa de grupos de trabalhos com diversos órgãos, agindo na "cola", juntando os instrumentos para promover o desenvolvimento industrial. Como exemplo foi citado o projeto da rede de smart grid.

Comentários Gerais do Painel:

- Aperfeiçoamento do marco legal voltado para inovação
- Empresas inovam pouco. Lei do Bem é pouco usada.
- Recursos de P&D tem que passar sempre por ICTs.
- Dar aos institutos privados e áreas de P&D das empresas o mesmo acesso a recursos que dá aos institutos públicos (no caso das empresas, exigindo a contrapartida em função do seu porte).
- Reduzir o tempo de concessão de financiamentos e subvenções pela FINEP e BNDES.

Um ano pode ser muito tempo quando falamos em inovação - a ideia ou melhoria pode não fazer mais sentido.

- Propriedade intelectual é algo muito crítico e baseado em relações de confiança entre os parceiros

- Dados e acesso a dados - instrumentos para interpretar os dados (analytics)

- "Software é tudo, mas é meio, a não ser para quem faz software básico. Inovação, no mundo inteiro e em todos os setores da economia, está sendo escrita em software, de robôs industriais a entregas especiais, passando pela padaria de Seu Manuel. Uma das pouquíssimas chances do Brasil inovar, em escala mundial e gerando trabalho e emprego sofisticado, bem remunerado e em escala é inovação com software."

2.5. PAINEL 5 – ESTRATÉGIA DE NUVEM/GOVERNO FEDERAL

Este painel foi composto pelos seguintes palestrantes: Bruno Pacheco – Coordenador de Modernização de Legados do SERPRO; Rodrigo Assumpção – Presidente DATAPREV; e Nazaré Bretas – Secretária Adjunta da SLTI - MPOG

Os principais posicionamentos dos membros do painel estão indicados a seguir por instituição:

SERPRO

Podemos resumir os pontos levantados pelo representante do Serpro. Ele cita os principais problemas tecnológicos do uso da computação em nuvem atuais e nos próximos 5 e 10 anos, resumidos da seguinte forma:

1. Integrar o Modelo de Nuvem com o Modelo real
2. Adaptar o modelo de gestão operacional real para o de Nuvem
3. Identificar arquiteturas de alto desempenho para a análise de dados
4. Disponibilizar serviços de Nuvem tais como plataforma como serviço e software como serviço
5. Estabelecer modelos de auto provisionamento de ambientes
6. Estabelecer modelo de transbordo entre nuvens heterogêneas

DATAPREV

O representante da Dataprev inicia sua fala dizendo que a TI determina os rumos do futuro e que seria relevante focar mais nos desafios do Brasil do que nos desafios da computação. Desafios tecnológicos são resolvidos ao longo do tempo. Novas interfaces de comunicação com as máquinas estão surgindo. Existe um grande desafio na TI pública e isso não pode ser um obstáculo nas dificuldades da gestão pública. Falta a infraestrutura

básica de dados. Maior desafio para o uso da computação em nuvem é a adaptação e as interfaces com os sistemas legados. A ausência de processos na informatização leva ao desenvolvimento de sistemas básicos e que não são necessariamente relevantes. Quando ele fala de governo não é só o federal, existe um número grande de municípios praticamente fora da discussão de TI e este é um problema de gestão. O modelo de licenciamento atual impede a interação com grandes players de nuvem por falta de um modelo de negócio adequado. Muito destes pontos estão ligados à questão de segurança.

SECRETARIA DE LOGÍSTICA E TECNOLOGIA DA INFORMAÇÃO DO MPOG

A representante da SLTI comenta que as responsabilidades da SLTI são: sistemas de logística e normas de contratação, normatização de transferências voluntárias e administração dos recursos de TI dos órgãos do Governo Federal. Atualmente trabalhando no código de CT&I para o MCTI.

Em TI existe uma grande fragmentação dos órgãos, que "compram" as mesmas coisas por conta da autonomia orçamentária. Esta divisão do orçamento é agravada pelo fato de historicamente as áreas de TI terceirizarem seus serviços. Esta foi uma das razões para a criação de um grande controle de TI e estruturação das áreas para pensar o orçamento de TI como um todo. Ações:

- constituição de uma infraestrutura de conectividade (Serpro): infovia Brasilia
- cadastro do conjunto dos servidores públicos do país (Dataprev)
- compras compartilhadas de bens.

O potencial emprego de computação em nuvem no Executivo Federal é o seguinte:

1. IaaS - Consolidação (gradual) de Nuvem Privada para uso de órgãos do SISP (órgão responsável pela administração de recursos de Tecnologia da Informação) – Infra Estruturas Críticas.

2. Paas - Seleção de serviços de uso em comum com baixa especialização por órgão: correio, agenda, gestão administrativa, "governança", projetos.

Também foi comentado sobre a adoção de TI Verde para redução geral de custos (de energia, de emissão de CO_2, de custos, entre outros).

2.6. PAINEL ESPECIAL SOBRE CENTROS DE P&D INSTALADOS NO BRASIL

Este painel especial traz a palavra de dirigentes de centros de P&D estrangeiros instalados no Brasil: Karin Breitman – General Manager do Centro de P&D da EMC; Claudio Pinhanez – Manager, Service Systems Research, do IBM Research; Fabio Tagnin – Diretor de expansão de mercados da Intel Brasil; Raimundo Nonato da Costa – Diretor Nacional de Tecnologia da Microsoft Brasil.

Este painel foi montando em consonância com um dos objetivos do plano TI Maior que é a atração de centros globais de P&D internacionais para o Brasil. Um dos pontos gerais discutidos foi que estes centros devem ser instalados nas diversas regiões do país (não apenas no Sudeste). Outro ponto fundamental para concretizar esta ação é a formação de pessoal qualificado em larga escala. Listamos a seguir os principais temas discutidos por cada um dos centros participantes do painel.

EMC

A apresentação foi centrada no surgimento de grande volume de dados desestruturados (na ordem de petabytes) e na tendência de geração de mais dados oriundos de redes sociais e dispositivos móveis. Em contrapartida constata-se a diminuição do custo do hardware. Uso de novas tecnologias nos próximos 5 anos: computação em nuvem, Big Data, hadoop, NoSQL, paralelização (map/reduce). Em resumo, ressalta-se a diminuição dos custos operacionais (computação em nuvem) e o uso de Big Data para administrar o crescimento dos dados.

IBM

Foram apontados os principais aspectos da computação a serem considerados no século 21, de acordo com grandes áreas: Ubiquidade: aumento da produtividade, internet das coisas, serviços por robôs; Biotecnologia: processamento genético, biologia sintética, programação de células; Analítica: análise de dados multimídia, sistemas cognitivos, modelagem de sistemas complexos; Social: bancos de Dados híbridos (estruturados e não estruturados), aplicações personalizadas, modelagem de sistemas sociais; Computação em Nuvem: segurança e privacidade, gerenciamento de sistemas legados, data-centers autônomos.

Intel

Foram apresentados a visão de inovação e os modelos de pesquisa da Intel. Os itens ressaltados nas pesquisas dos Intel Labs foram: circuitos, emulação e validação física, energia e sustentabilidade, micro arquitetura, computação paralela, embarcados e novos dispositivos. Os focos dos centros de pesquisa citados foram relativos à: computação em nuvem, computação social, computação de Big Data, computação visual, computação pervasiva e computação segura. Também foram apresentadas as principais invenções da Intel, um pouco da sua história e as principais tendências do futuro: miniaturização, novas tecnologias associadas aos sentidos humanos (voz, toque, entre outros), sistemas on-chip, dispositivos ligados à nuvem, hipercomputação.

Microsoft

O palestrante apresentou sobre a presença da Microsoft em várias regiões do Brasil e suas ações alinhadas ao plano TI Maior. As áreas de foco do Laboratório de Tecnologia Avançada instalado no Rio de Janeiro são Classificação de Documentos, Web Services e Pesquisa e Recuperação de Informação.

Comentários Gerais do Painel:
- Propriedade Intelectual no Brasil é complicado
- Operacionalização da integração dos laboratórios com a academia

2.7 CONSIDERAÇÕES SOBRE OS PRINCIPAIS TEMAS DISCUTIDOS

Muitos foram os temas da Computação identificados nas apresentações e discussões dos painéis que ocorreram durante o evento.

Achamos interessante relacionar o que foi discutido nos painéis com os Grandes Desafios da Computação definidos em sua primeira edição em 2006, por considerarmos esta edição como uma referência maior. Esta análise nos mostra que os Grandes Desafios da Computação definidos em 2006 ainda estão presentes hoje e, em especial, nos temas destacados das áreas estratégicas abordadas neste seminário: sistema bancário e financeiro, saúde, petróleo & gás, energia e defesa cibernética. É importante ressaltar que, apesar dos grandes desafios estarem ainda vigentes, surgiram novas tecnologias, a exemplo de computação em nuvem e as tecnologias relacionadas a Big Data, além de novos problemas e aplicações como é o caso das Cidades Inteligentes.

3. TERCEIRO SEMINÁRIO GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL – FASE 2

O 3º Seminário Grandes Desafios da Computação no Brasil – Fase 2 teve como objetivo promover redes de colaboração temáticas em função de problemas reais que envolvessem governo, indústria e academia. Para isso, através de uma chamada de trabalhos, o evento buscou identificar parcerias possíveis ou já existentes entre governo-indústria-academia, dentro do contexto dos grandes desafios da Computação destacados pela Sociedade Brasileira de Computação nos domínios de Sistema Bancário/Financeiro, Petróleo, Energia, Defesa Cibernética, Saúde e Educação, e Mobilidade. São eles:

- Gestão da Informação em grandes volumes de dados multimídia distribuídos;
- Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem natureza;
- Impactos em TICs na transição do silício para novas tecnologias;
- Acesso participativo e universal do cidadão brasileiro ao conhecimento;
- Desenvolvimento tecnológico de qualidade.

No intuito de aprofundar essas discussões, os participantes se reuniram em workshops temáticos divididos entre os domínios anteriormente mencionados. Também foi realizada uma rodada de negócios entre os participantes, a fim de que fossem identificados interesses em comum que pudessem dar início a projetos em parceria. A programação do evento foi completada com palestras, uma mesa de debate sobre os Desafios Computacionais no Governo e um painel que debateu as Empresas e os Desafios da Computação. Ao contrário de encontros anteriores, este seminário teve um perfil mais aberto a estudantes, professores e à própria iniciativa privada consumidora de tecnologia.

Na busca por parcerias entre governo, indústria e academia, a participação do Ministério de Ciência, Tecnologia e Inovação (MCTI) no evento foi essencial. O órgão possui um cadastro de empresas interessadas em aplicar parte de seus recursos em P&D em troca de isenções fiscais e a relação das instituições de ensino e pesquisa, institutos e incubadoras autorizadas a receber esses investimentos e desenvolver as soluções por meio da Lei de Informática.

"Essa é outra proposta do Seminário: promover a aproximação das empresas, entidades e instituições dos dois cadastros do MCTI para que o desenvolvimento de novas tecnologias possa ocorrer com mais facilidade", explica a coordenadora do Seminário, Claudia Motta.

Este relatório visa apontar as diretrizes das pesquisas em Ciência da Computação relacionadas aos domínios destacados. com base nas discussões realizadas durante o Seminário. Parte dessas discussões também integrará a programação do Congresso da SBC, a ser realizado no Recife, entre 20 e 23 de julho de 2015.

3.1 CHAMADA DE TRABALHOS

Através site do evento[10], foram recebidos 35 artigos de pesquisadores de todo o Brasil. A submissão dos artigos foi realizada por meio de um formulário online. Todos os artigos foram avaliados por dois revisores. Os artigos aceitos, para a apresentação principal e para os workshops temáticos, foram aqueles que os dois revisores deram aceite para apresentação. Os artigos premiados foram, dentre estes, os que tiveram as maiores

[10]http://www.sbcgrandesdesafios.nce.ufrj.br/

notas em cada domínio. Além disso, os artigos que tiveram pelo menos uma indicação para a rodada de negócios foram selecionados para esta sessão.

TABELA 1 – Quantidade de trabalhos enviados por tema

Tópico	Aceito	Rejeitado	Total válido *
Gestão da Informação em grandes volumes de dados multimídia distribuídos	9	2	11
Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem natureza	4	3	7
Impactos para a área da Computação da transição do silício para novas tecnologias	2	3	5
Acesso participativo e universal do cidadão brasileiro ao conhecimento	1	5	6
Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos	4	1	5

A premiação dos artigos ocorreu no dia 18 de setembro no Auditório do Roxinho. Os artigos premiados e suas respectivas categorias estão descritas abaixo:

- Primeiro lugar geral do evento: "Ciência de Dados". Autores: FábioPorto e Artur Ziviani – Laboratório Nacional de Computação Científica.

- Primeiro lugar na categoria Saúde: "Sistema de Informação em Saúde Silvestre – SISS-Geo". Autores: Marcia Chame, Helio J. C. Barbosa, Luiz Gadelha, Douglas A. Augusto, Eduardo Krempser, Livia Abdalla – Fundação Oswaldo Cruz e Laboratório Nacional de Computação Científica.

- Primeiro lugar na categoria Mobilidade: "Descoberta de Padrões em Sistemas Urbanos Complexos". Autora: Ana Lucia Bazzan – Universidade Federal do Rio Grande do Sul.

- Primeiro lugar na categoria Educação: "Desafios e Oportunidades em Neurociência Computacional na Educação Brasileira". Autores: Raimundo José Macário Costa, Luís Alfredo Vidal de Carvalho, Emilio Sánchez Miguel, Renata Mousinho, Renato Cerceau, Lizete Pontes Macário Costa, Sérgio Manuel Serra da Cruz– Universidade Federal Rural Rio de Janeiro, Universidade Federal Rio de Janeiro, Universidad de Salamanca, Universidade do Estado do Rio de Janeiro, Agência Nacional de Saúde Suplementar, Programa de Educação Tutorial.

- Primeiro lugar na categoria Sistema Bancário/Financeiro: "Contribuições e Desafios de Tecnologias de Dados Sociais na Expansão e Melhoria do Sistema Bancário e Financeiro no Brasil: Eficiência, Produtividade e Inclusão Social". Autor: Claudio Pinhanez – IBM Research.

- Primeiro lugar na categoria Petróleo/Energia: "Interoperabilidade Semântica na Cadeia de Exploração de Petróleo". Autores: Mara Abel, Luis Fernando De Ros, Joel Carbonera, Sandro Rama Fiorini, Alexandre Lorenzatti – Universidade Federal do Rio Grande do Sul.

Os prêmios e certificados foram entregues aos vencedores pelos membros da mesa de abertura (descrita em seguida). Como parte da programação do evento, os autores dos artigos premiados tiveram a oportunidade de apresentá-los para todos os participantes.

Os artigos na íntegra encontram-se no Anexo.

4. ABERTURA

Na mesa de abertura do evento estavam presentes:

- O Magnífico Reitor da Universidade Federal do Rio de Janeiro, professor Carlos Antonio Levi da Conceição.

- O Secretário de Política de Informática do Ministério da Ciência, Tecnologia e Inovação, professor Virgílio Almeida.

- O Presidente da Sociedade Brasileira de Computação, professor Paulo Roberto Freire Cunha. A professora da Universidade Federal do Estado do Rio de Janeiro e Coordenadora Geral do Seminário Grandes Desafios da Computação, Flávia Maria Santoro.

- A Diretora do Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais e Coordenadora Geral do Seminário Grandes Desafios da Computação, Claudia Lage Rebello da Motta.

Claudia Motta iniciou sua fala agradecendo pela presença de todos e principalmente dos membros da mesa de abertura. Claudia Motta falou brevemente sobre a trajetória dos Seminários Grandes Desafios da Computação e reforçou a importância de se identificar no evento as possibilidades de parceria entre Academia, Setor Privado e Governo e conseguir, através do debate, estabelecer uma ponte entre eles.

Em seguida, a professora Flávia Santoro assumiu a palavra e falou sobre a programação do evento como um todo, com destaque para a seleção dos artigos que seriam premiados e sua relação com os temas identificados na Fase 1 do Seminário, além

de explicar o funcionamento dos workshops temáticos e o networking de projetos.

O Presidente da Sociedade Brasileira de Computação, professor Paulo Roberto Freire Cunha, discursou sobre a importância da aproximação da SBC com as empresas e o governo, para a qual a instituição vem trabalhando através da Diretoria de Articulação com Empresas, e sobre a relação desta necessidade de aproximação com a organização dos Seminários Grandes Desafios da Computação.

O Professor Virgílio Almeida deu prosseguimento às falas de abertura ressaltando a importância de iniciativas como este seminário para o nosso país, especialmente tendo em vista a relevância do setor de Tecnologias da Informação e Comunicação para o governo brasileiro devido a sua representatividade dentro de nossa economia (aproximadamente 7% do PIB do país) e a sua capacidade de gerar inovação. O Secretário destacou ainda a importância do desenvolvimento de tecnologia própria dentro do país, o que pode ser facilitado através da Lei de Informática.

Por fim, O Magnífico Reitor da Universidade Federal do Rio de Janeiro discursou sobre o surgimento de desafios cada vez mais complexos na área da computação, destacando os avanços estratégicos das TICs e a produção de resultados cada vez mais relevantes em todas as esferas da sociedade moderna. Segundo o Reitor, o setor se desenvolve numa velocidade que exige muito dos pesquisadores e da própria universidade, o que exige um esforço constante que torne possível internalizar esses avanços. Ele reforçou ainda a visão do secretário de que os avanços nesse setor impulsionam avanços em todos os outros setores que compõe nossa economia, destacando a importância de um evento como o Seminário.

5. GRANDE DESAFIO BRASILEIRO: PRODUTIVIDADE. E O QUE É QUE A CIÊNCIA, TECNOLOGIA E INOVAÇÃO TÊM A VER COM ISSO?

O professor Silvio Meira, da Universidade Federal de Pernambuco, iniciou sua palestra falando sobre a enorme mudança infra estrutural ocorrida na última década e o impacto dessa mudança na economia global. Dentro deste contexto, o crescimento de todos os países do mundo é cada vez maior, porém desigual. Segundo os dados apresentados, enquanto outros países como EUA e China crescem até 100% em uma década, o Brasil cresce apenas 40%.

Silvio apontou que o principal problema do Brasil seria a produtividade e que o condicionante para que esta produtividade aumente significativamente é a Educação. Nos últimos 50 anos o Brasil não teria evoluído nada no setor de serviços, ou seja, a distância entre o Brasil e os outros países só aumenta cada vez mais. Além disso, em função da baixa produtividade brasileira, grandes centros de pesquisa são instalados somente no exterior.

O professor pontuou que o Brasil seria uma grande complicação, e deu como exemplo a questão dos impostos: uma empresa no Brasil gasta em média 2.600 horas por ano para pagar seus impostos, enquanto que uma empresa no Chile gasta apenas 396. O problema não seria a carga tributária em si, mas sim a pouca produtividade que leva a um altíssimo gasto de horas para pagá-la.

O pesquisador colocou que o Estado deveria deixar de interferir, pois temos um Estado que se envolve em tudo e que acabaria, consequentemente, não fazendo nada direito, contribuindo, assim, para um atraso no desenvolvimento de inovações. Nas palavras do palestrante: "Inovar é resolver NOVOS problemas de forma que ninguém nunca antes atacou".

Uma das questões que mais impede que a inovação ocorra é que CEOs de grandes empresas, governantes e indivíduos em cargos de liderança tendem a acreditar que ao invés de inovar é sempre melhor fazer o que já está dando certo. Ter-se-ia, assim, alguns paradoxos da inovação:

1) Inflexibilidade – Temos em nossos ambientes de trabalho estruturas, controle e modelos, e não liberdade, flexibilidade e emergência.

2) Falha – Ninguém aprende com o sucesso, mas só se premia o sucesso. Quanto mais sucesso tal estratégia tem, mais se fortalece a crença de que ela deve ser mantida. A tendência é de manter os recursos para aqueles que já fazem sucesso, e não investir em coisas novas.

3) Conhecimento – Quanto mais eu sei de maneira fechada, menos aberto eu fico e menos eu me interesso por outras áreas. Quanto mais profundamente eu entendo do meu negócio, menos eu quero saber do negócio dos outros.

4) Alinhamento – Quanto mais gente está indo na mesma direção, mas difícil e arriscado fica inovar e chamar atenção para outras direções.

Silvio mostrou ainda que em todas as ondas inovação (revoluções de informação e tecnologia) o Brasil chegou completamente atrasado e se tornou um simples consumidor.

Sendo assim, o que pesquisador sugere que poderia ser feito, do ponto de vista criativo e inovador, para nos tornarmos competitivos no mercado mundial é criar e desenvolver tecnologia, e não apenas importá-la do exterior. Boa parte da pesquisa brasileira seria completamente inútil e sem importância, pois muitos pesquisadores desenvolvem tecnologia que não tem qualquer demanda no mercado.

O professor conclui dizendo que o maior desafio dos próximos anos seria alinhar importância e relevância, para não desperdiçar tempo e recursos desenvolvendo tecnologia e conhecimento "inútil" e passar a investir no desenvolvimento de tecnologia própria que nos torne competitivos no mercado internacional.

6. MESA DE DEBATE

A mesa Desafios Computacionais no Governo foi composta pelo Secretário de Política de Informática do Ministério da Ciência, Tecnologia e Inovação, professor Virgílio Almeida, pelo senhor Adalberto Afonso Barbosa, também do Ministério da Ciência, Tecnologia e Inovação, e pela Pró-reitora de Pós-Graduação e Pesquisa da Universidade Federal do Rio de Janeiro, professora Debora Foguel. O Presidente da Sociedade Brasileira de Computação, professor Paulo Roberto Freire Cunha, se encarregou da moderação.

6.1 APRESENTAÇÃO DO PROFESSOR VIRGÍLIO ALMEIDA

O debate teve início com a apresentação do Secretário Virgílio Almeida, que apresentou um panorama das Tecnologias da Informação e da Comunicação no Brasil. Segundo os dados apresentados, o Brasil está entre os três maiores mercados em sites de mídias sociais no mundo, o que mostra que há uma grande abertura da sociedade brasileira às novas tecnologias.

Também foram apresentados dados que mostravam a importância do setor digital para o país e casos de sucesso do uso de tecnologias digitais em outros setores, como: setor de RH; sistema financeiro; governo eletrônico; energia, petróleo e gás; agricultura e manufatura. Com isso, o secretário destacou que a transversalidade seria a chave para a inovação dos vários setores.

Em seguida Virgílio falou sobre os blocos fundamentais da política de TICs no Brasil. São eles:

- Hardware, semicondutores e displays
- Software
- Ciber-Infraestrutura para P&D

Foi destacado ainda o importante papel da Lei de Informática[11] dento deste contexto de desenvolvimento. Inicialmente o objetivo da Lei era fomentar a manufatura no país através deste incentivo, e o próximo passo seria pensar estrategicamente o projeto do país e caminhar para uma integração de políticas públicas.

Como conclusão, o Secretário falou sobre sua meta para o futuro, o Brasil Digital, que consistiria nos seguintes pontos:

1) Natureza transversal das TICs transformando a sociedade
2) Softwares sendo capazes de abranger cada vez mais a economia do país
3) Visão de longo prazo que de ênfase a qualificação de pessoas

[11]http://www.mct.gov.br/index.php/content/view/2189.html

Mas, para isso, é preciso que o Brasil evolua e se torne um país onde seja possível encontrar acesso à internet para todos, cidadãos habilitados, um setor da economia da informação forte, o uso das TICs por parte de todas as empresas e onde haja crescimento sustentável e inclusivo.

6.2 APRESENTAÇÃO ADALBERTO AFONSO BARBOSA

O senhor Adalberto falou um pouco mais sobre a Lei de Informática e pontuou que parte da renúncia fiscal possibilitada através da Lei deve ser investida em convênio com algum tipo de instituição de ensino ou pesquisa. De acordo com os dados apresentados, isso seria equivalente a aproximadamente 600 milhões de reais. Foi destacado ainda que a Lei de Informática possui uma grande abrangência, o que permite que as contrapartidas se estendam para outros tipos de setores, como: automação industrial, serviços, comercial e hospitalar. Em seguida Adalberto apresentou os resultados da pesquisa realizada com empresas em relação aos Grandes Desafios da Computação.

Resultado Pesquisa de Interesse sobre os Grandes Desafios da Computação

Mensagens eletrônicas foram enviadas para membros de diferentes empresas integrantes dos cadastros da MCTI e da SBC com a solicitação de que um breve questionário fosse respondido no site do evento. Cada respondente era convidado a apontar os interesses específicos de sua empresa em relação a 62 itens divididos em de 5 grandes áreas:

1) Gestão da Informação em grandes volumes de dados multimídia distribuídos;

2) Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem natureza;

3) Impactos em TICs na transição do silício para novas tecnologias;

4) Acesso participativo e universal do cidadão brasileiro ao conhecimento e

5) Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos

O questionário permitia múltipla escolha, o que gerou um total de 2058 respostas de 137 respondentes. Após o cruzamento dos dados adquiridos, os resultados gerais da pesquisa apontam que os itens destacados como mais relevantes por diferentes representantes do setor privado no Brasil são (por ordem de relevância), apresentados na tabela 2.

TABELA 2 – Resultados gerais da pesquisa	
Itens	**Área**
Sistemas de alta confiabilidade	5
Dispositivos embarcados e móveis	3
Internet das coisas	5
Dispositivos inteligentes (smart meters, smart grid)	3
Componentes digitais estratégicos (semicondutores, displays, OLED, grafeno)	3
Business Intelligence (Sistemas analíticos e de predição, Big Data)	1
Interface Homem-máquina	4
Programação em pequenos dispositivos	5
Sensoriamento remoto	2
Cidades Inteligentes	1
Reconhecimento de padrões	2
Processamento de Imagens	2
Sistemas analíticos e de predição	1
Ética	4
Consideração do lado humano no desenvolvimento dos sistemas	4

7. WORKSHOPS TEMÁTICOS

Cada workshop temático contou com a apresentação de um ou mais trabalhos, relacionados aos temas que seriam debatidos, e com a presença de um facilitador que, em seguida, se encarregaria de apresentar para todos os participantes do evento um resumo do que havia sido discutido e um resumo dos desafios identificados.

7.1 DEFESA CIBERNÉTICA

Facilitadora: Professora Renata Galante

Artigo Apresentado:

1) Segurança Cibernética: Alavanca Estratégica de Software e Serviços de Tecnologia da Informação e da Comunicação.

Autores: Ricardo Dahab (UNICAMP) e Michele Nogueira (UFPR)

7.1.1 DEBATE:

CLOUD COMPUTING

O ambiente de computação em nuvem (CC) está sujeito a invasões criminosas que necessitam de coleta online de evidências. A versão comercial (business-driven) de CC não permite acesso além da Virtual Machine (VM) oferecida para o processamento. Uma maneira de viabilizar a coleta de evidências sem alterar a cena do crime cibernético é desenvolver técnicas que permitam a cópia de memória, a partir do lado externo a VM.

TRANSPARÊNCIA

A continuidade da comunicação dos dados e a sua disponibilidade ao cidadão para o lazer, saúde, educação, mobilidade urbana, bem como ao governo para diferentes ações de proteção nacional demandam ações que garantam o acesso e a migração a diferentes tipos de redes sem fio estruturadas e não estruturadas e sua integração com redes cabeadas, visto que a cada tipo de rede apresenta restrições de recursos e de escalabilidade. Além disso, essa transparência de continuidade se torna essencial por conta do grande volume de dados que teremos ainda mais no futuro. Logo, a comunicação estará sendo prejudicada por várias formas de ataques programados e não programados que exigirá correções em tempo real e sem prejuízo perceptível ao cidadão, empresas e estruturas do governo.

INTERNET OF THINGS (IoT)

Os dispositivos de internet das Coisas (IoT) estarão cada vez mais presentes no nosso dia a dia, mas estão expostos a ataques como qualquer outro dispositivo que tem um IP. É necessário que sejam desenvolvidas técnicas de detecção de intrusão e proteção de privacidade energeticamente eficientes.

DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

As concessionárias de distribuição de energia elétrica (CDE) são fraudadas com roubos importantes de energia elétrica todos os dias. As redes elétricas inteligentes (Smart Grid) vão permitir que a CDE tenha acesso ao medidor inteligente (Smart Meter) em tempo real. Porém, como é possível descobrir onde o roubo está acontecendo, dado que há de considerar também as perdas técnicas?

CURVAS ELÍPTICAS

Seleção de curvas elípticas para assinaturas e confidencialidade. Os padrões existentes (NIST, Brainpool) apresentam indícios de manipulação, ou de manipulabilidade, por agências estrangeiras de inteligência. Construir curvas robustas de maneira a evitar esses cenários, mas preservando eficiência em todas as métricas (como alto desempenho e pequeno consumo de energia), é essencial para garantir a fidedignidade da infraestrutura nacional de chaves públicas.

ALTERNATIVAS À CRIPTOGRAFIA CONVENCIONAL DE CHAVE PÚBLICA

A necessidade de substituir chaves RSA por esquemas mais eficientes é ilustrada pelo uso crescente de criptografia elíptica, mas essa tecnologia é notoriamente suscetível a computadores quânticos, que, embora não sejam ainda uma realidade tecnológica, são uma preocupação para estabelecer segurança a longo prazo. Esquemas clássicos resistentes a ataques quânticos são possíveis, mas só recentemente vêm atraindo investigações de potencial tecnológico. Isso torna o assunto de interesse estratégico, em especial nos aspectos de migração e manutenção/extensão das funcionalidades clássicas para esse ambiente pós-quântico.

7.1.2 DESAFIOS IDENTIFICADOS:

1) Necessidade de desenvolvimento de soluções práticas para garantia de privacidade para bases de dados de saúde armazenados em nuvens públicas que possibilitem pesquisas (mesmo que básicas) nos dados cifrados.

2) Necessidade de desenvolvimento de técnicas de auditoria setorizada da rede de distribuição e do consumos nos medidores inteligente poderiam dar indícios de onde a fraude está ocorrendo.

3) Construir curvas robustas é essencial para garantir a fidedignidade da infraestrutura nacional de chaves públicas.

4) Necessidade de substituir chaves RSA por esquemas mais eficientes.

5) Autenticação de documentos online: falsificações ocorrem o tempo todo para denegrir a imagem de terceiros, falsificar acontecimentos, etc.

6) Atribuição de fonte para documentos divulgados na web: descobrir quem gerou o que, de onde veio, para fazer rastreamento e atribuir culpabilidade.

7.2 EDUCAÇÃO

Facilitador: Professor Marcelo Duduchi

Artigos Apresentados:

1) Desafios e Oportunidades em Neurociência Computacional na Educação Brasileira.

Autor: Raimundo José Macário Costa, Luís Alfredo Vidal de Carvalho; Emílio Sánchez Miguel, Renata Mousinho, Renato Rercear, Lizete Pontes Macário Costa, Sérgio Manuel Serra da Cruz, Universidade Federal Rural do Rio de Janeiro.

2) Participação popular e tecnologias: experiências e desafios

Autores: Cristiano Maciel (UFMT), Claudia Cappelli (UNIRIO), Cleyton Slaviero (UFF)

3) Sistemas Educacionais Inteligentes

Autores: Carlo Oliveira (UFRJ), Carla Verônica Marques (UFRJ), Claudia L R Motta (UFRJ), Maira Froes (UFRJ), Priscila Lima (UFRJ), José Otávio Silva (UFRJ)

7.2.1 DEBATE

O Workshop contou com a participação de aproximadamente 40 integrantes, em sua maioria professores e pesquisadores. A pequena participação de representantes de empresas e governo nas discussões fez com que fosse sugerida uma revisão do modelo do evento como um todo:

a) Alinhar de maneira objetiva os desafios aos interesses da empresa e governo;

b) Identificar que tipos de pesquisas são realmente necessárias para que haja investimento;

c) Ir até a empresa e governo com uma proposta concisa e bem definida.

7.2.2 DESAFIOS IDENTIFICADOS

1) Usar a Neurociência Computacional para identificar precocemente os transtornos de aprendizagem a partir de reconhecimento de padrões;

2) Desenvolver e aplicar sistemas inteligentes educacionais para a educação baseada e promovida por princípios na interação adaptativa à assinatura cognitiva do aluno;

3) Buscar a invenção da criança cidadã para que se possa promover a e-democracia e a e-participação no Brasil.

4) Desenvolvimento de sistemas robustos, fáceis de operar com informação transparente que permitam o acesso à informação por processos consultivos e conduzam à participação do cidadão na tomada de decisão.

7.3 MOBILIDADE/OUTROS

Facilitador: Professora Ana Carolina Salgado

Artigos Apresentados:

1) Concepção de Sistemas Integrados Tolerantes a Efeitos de Radiação

Autores: Ricardo Reis (UFRGS), Fernanda Kastensmidt (UFRGS)

2) Processamento de Redes Dependentes do Tempo de Larga Escala

Autores: José Antonio Macêdo (UFC), Marco Antonio Casanova (PUC-Rio), Regis Magalhães (UFC), Chiara Renso (KDDLAB ISTI/CNR - Italy), Raffaele Perego (ISTI - CNR - Italy), Regis Melo (Sagarana Tech)

3) Redes de Sensoriamento Participativo

Autores: Antonio Alfredo Ferreira Loureiro (UFMG), Felipe França (COPPE-UFRJ), Priscila M V Lima (UFRJ), Leonardo Oliveira (UFMG), Pedro Olmo Vaz de Melo (UFMG), Thiago Silva (UFMG), Olga Goussevskaia (UFMG), Italo Cunha (UFMG)

4) Enriquecimento Semântico, Análise e Mineração de Dados sobre Movimento com Ontologias e Dados Ligados

Autores: Vania Bogorny (UFSC), José Antonio Macêdo (UFC), Chiara Renso (KDDLAB ISTI/CNR - Italy), Alessandra Raffaeta (Ca Foscari University - Italy), Yanis Theodoridis (UPRC - Greece), Nikos Pelekis (UPRC - Greece)

7.3.1 DEBATE

O workshop contou com a presença de 12 pessoas e a apresentação três trabalhos, além do artigo vencedor do primeiro lugar na categoria Mobilidade do Seminário, que já havia sido apresentado. Todos os trabalhos apresentados estavam preocupados com grandes volumes de dados urbanos e eram bastante complementares. O primeiro trabalho tratava da variação temporal nas sugestões de rota de trânsito com atualização em tempo real. O segundo trabalho falava de redes de sensoriamento e sensoriamento participativo. O terceiro trabalho abordou o enriquecimento semântico, as trajetórias semânticas estruturadas, a identificação de padrões e a mineração dos dados.

Após as apresentações, os participantes debateram sobre algumas das dificuldades da pesquisa na área de mobilidade. Essas pesquisas seriam realizadas por pessoas de áreas diferentes, o que mostra que se trata de um tema transdisciplinar. No entanto, não há uma ação da SBC para temas que enquadre esse tipo de temas transversais. Portanto, os participantes sugeriram que o tema debatido no workshop temático fosse enquadrado como CIDADES INTELIGENTES, e não Mobilidade.

Além disso, foi debatido também o problema do dado público. Os sensores, câmeras e GPS espalhados pela cidade não são públicos, e, apesar de não serem disponibilizados para as universidades, são vendidos para grandes empresas como o Google. Sendo assim, os pesquisadores ficariam à mercê das empresas que coletam esses dados, e gostariam de pensar em ações que possam ser feitas para mudar essa conjuntura.

7.3.2 DESAFIOS IDENTIFICADOS

1) Descoberta de padrões
2) Estatísticas dos dados
3) Integração de dados
4) Localização e rastreamento
5) Contexto – informações contextuais gerais
6) Segurança e privacidade
7) Gerenciamento de energia
8) Uso de Cloud

9) Trajetórias com informações semânticas

10) Dados abertos e ligados

11) Mídias sociais

12) Enriquecimento semântico

13) Mineração de dados

14) Gerenciamento grande volume de dados

15) Variação temporal

16) Atualização em tempo real

7.4 PETRÓLEO E ENERGIA

Facilitador: Professor José Viterbo

Artigos Apresentados:

1) Monitoramento e adaptação de transformações em dados científicos ao longo de execuções paralelas em ambientes de processamento de alto desempenho

Autores: Marta Mattoso (COPPE/UFRJ) e Daniel de Oliveira (UFF)

2) Desenvolvimento de Sistemas de Software para Aumento da Segurança na Cadeia de Mineração

Autores: Cleidson de Souza (UFPA), Schubert Carvalho (EPFL – Switzerland) e Gustavo Pessin (USP)

7.4.1 DEBATE

O Workshop Temático sobre Petróleo e Energia contou com a presença de 17 participantes.

O primeiro trabalho apresentado, "Monitoramento e adaptação de transformações em dados científicos ao longo de execuções paralelas em ambientes de processamento de alto desempenho" destacou como principal desafio a capacidade de extrair dados de proveniência relacionados aos dados de domínio, o que permitiria rastrear e reproduzir as transformações realizadas ao longo do processo científico. Nesse escopo, o apresentador identificou ainda dois desafios relacionados: a visualização de dados intermediários e produção de ontologias capazes de representar o domínio dos dados de interesse.

O segundo trabalho, "Desenvolvimento de Sistemas de Software para Aumento da Segurança na Cadeia de Mineração" destacou como principal desafio o desenvolvimento de tecnologias de informação e comunicação (TIC) no contexto da Internet das Coisas (IoT), com foco na segurança da população. Este desafio pode ser decomposto em diversos fatores, tais como, questões de engenharia de software relacionados ao de-

senvolvimento de aplicações para IoT, boas práticas para o projeto de sensores para sistemas de alto risco, e a formação de recursos humanos para estes contextos. O apresentador destacou os aspectos que diferenciam o cenário de IoT de cenários tradicionais, que são a heterogeneidade e geração contínua de dados. Aliado ao grande número de fontes de dados, estes aspectos exigem a aplicação de técnicas de Big Data nos cenários descritos.

Ao final, os participantes do evento concordaram que seria uma boa ideia a criação de uma comunidade virtual para continuar a discussão iniciada pelo grupo naquele encontro, chegando a uma lista de resultados. A discussão dos resultados com a comunidade em geral seria realizada no CSBC 2015. Além disso, os participantes sugeriram que fossem propostas sessões específicas nos diversos congressos temáticos da SBC (SBBD, SBRC, SBES, IHC, etc) para a discussão de desafios aplicáveis a domínios específicos da computação.

7.4.2 DESAFIOS IDENTIFICADOS

1) Capacidade de extrair dados de proveniência relacionados aos dados de domínio.

2) Visualização de dados intermediários.

3) Produção de ontologias capazes de representar o domínio dos dados de interesse.

4) Desenvolvimento de tecnologias de informação e comunicação (TIC) no contexto da Internet das Coisas (IoT), com foco na segurança da população.

5) Necessidade do desenvolvimento da visualização de dados não apenas voltada para a proveniência de dados, mas visando também o compartilhamento de bons resultados que possam ser obtidos em processos determinísticos influenciados por grandes números de variáveis.

6) Importância de associar problemas sociais práticos ao ensino de técnicas básicas e ao conteúdo das disciplinas na formação de recursos humanos de computação.

7) Grande dificuldade na obtenção de dados reais para o apoio ao ensino e à pesquisa.

7.5 SAÚDE

Facilitador: Professor Carlos Ferraz

Artigos Apresentados:

1) Biometria por Padrões Papiloscópicos, Expressões Faciais e Gestos

Autores: Olga Bellon (UFPR)

2) Otimização de Componentes em Sistemas Integrados Visando Reduzir o Consumo de Energia

Autores: Ricardo Reis (UFRGS)

3) Deduplicação de Dados em Larga Escala em Paralelo no Domínio de Saúde

Autores: Carlos Eduardo Santos Pires (UFCG), Dimas C. Nascimento (UFCG), Demetrio Mestre (UFGC)

7.5.1 DESAFIOS IDENTIFICADOS

1) Dificuldade de coleta de dados (das impressões digitais de neonatais, principalmente).

2) Baixo consumo e mais confiabilidade para dispositivos de saúde implantáveis (ex. Marcapasso) e de monitoramento externo.

3) Chips tolerantes a radiação para redução de impacto em aplicações de saúde.

4) Grande volume e heterogeneidade de dados.

5) Regras de contexto (modelagem) para previsões melhores.

6) Uso de dispositivos móveis para ampliar captura de dados (problemas de rede e energia para os dispositivos).

7.6 SISTEMA FINANCEIRO / BANCÁRIO

Facilitador: Professor Altigran Soares da Silva

Artigos Apresentados:

1) Big Data, Little Data e Better Data em Sistemas de Recomendação

Autores: Priscila Lima, Claudia L. R. da Motta, Adriano J. O. Cruz, Antonio J. Alencar, Eber A. Schmitz, Jonas Knopman, Cabral Lima (UFRJ).

7.6.1 DEBATE

O Workshop contou com aproximadamente 12 participantes, entre eles professores e pesquisadores, alunos de graduação e pós-graduação e profissionais (IBM e Banco Central). Nenhum profissional de instituição bancária de varejo ou do mercado financeiro esteve presente. Foi apresentado um artigo, "Big Data, Little Data e Better Data em Sistemas de Recomendação", por Priscila Lima (UFRJ).

Em seguida, foi aberta a discussão em torno de quais seriam as demandas do setor para a pesquisa em Ciência da Computação. Foram então mencionados e brevemente debatidos vários tópicos, sendo estes tópicos organizados em duas categorias: B2B (Relação entre Instituições) e B2C (Relações com Clientes).

De maneira transversal, foram também mencionadas questões como dinheiro eletrônico e sistemas de pontuação com forma de pagamento.

Ao final, os participantes concluíram que mesmo com essa breve discussão já foi possível identificar demandas de muita importância para o setor, com as quais a pesquisa em Ciência da Computação poderia efetivamente contribuir. Assim, ficou a sugestão de que o tema fosse aprofundado, de preferência com o envolvimento de profissionais de mercado.

7.6.2 DESAFIOS IDENTIFICADOS

Relação entre Instituições (B2B)

1) Integração de Dados

2) BI on-line

3) Algorithmic Trading

4) Business Compliance

Relações com Clientes (B2C)

1) Eficiência: diminuição de custos, modelos preditivos de riscos, redução de custos de infraestrutura.

2) Produtividade: informações melhores para gerentes de conta, tornar as agências mais eficazes e eficientes.

3) Satisfação do cliente: recomendação de produtos financeiros, melhoria e redução de procedimentos burocráticos e de segurança, redução de custos transacionais e de serviços.

4) Inclusão social: aumentar a parcela da população que participa do sistema bancário/financeiro, assistir adequadamente idosos e portadores de deficiências, fomentar a educação financeira.

8. PALESTRA: REDE INOVAPUC, PREPARANDO UM AMBIENTE PARA INTERAÇÃO UNIVERSIDADE – EMPRESA – GOVERNO NA PRÁTICA

Em sua palestra, o professor Avelizo Zorzo apresentou a Rede de Inovação e Empreendedorismo da Pontifícia Universidade Católica do Rio Grande do Sul, Rede InovaPUCRS[12]. Trata-se de uma rede de inovação que envolve desde o aluno de graduação até o Reitor da universidade.

Avelino destacou a mudança de posicionamento da Universidade com o intuito de atrair os melhores alunos possíveis, gerando, dessa forma, um maior capital intelectual que

[12]http://www3.pucrs.br/portal/page/portal/inovapucrs/Capa/Institucional

possa ser revertido efetivamente para a sociedade. Como resultado dessa medida tem-se um "ciclo virtuoso":

Bons alunos ingressam na universidade; Capital intelectual é gerado; O capital intelectual gerado é consumido pelas grandes empresas; Bons alunos são atraídos para a universidade.

O processo de inovação teria começado na década de 90. Um dos programas mais importantes nos quais a universidade investiu foi o programa 1000 em 2000, que consistia no investimento na formação dos professores da própria PUC – A meta era chegar ao ano 2000 com 1000 mestres e doutores no quadro de profissionais. Além disso, houve um forte investimento em infraestrutura na universidade como um todo e a elaboração de um planejamento estratégico focado no futuro.

Voltando à Rede InovaPUC, Avelino explica que sua missão seria congregar o conjunto de atores, ações e mecanismos para fomento do processo de inovação e empreendedorismo da PUCRS. Para tanto, a Rede procura articular todos os envolvidos na tríade "ensino, pesquisa e extensão", sendo formada por distintas unidades periféricas com objetivos específicos bem claros. São elas:

NÚCLEO EMPREENDEDOR – estimula o empreendedorismo, organiza uma série de eventos.

INSTITUTO DE PESQUISA E DESENVOLVIMENTO – ajuda o aluno a desenvolver tecnicamente o produto pensado no Núcleo Empreendedor.

INCUBADORA RAIAR – após a pesquisa ter sido desenvolvida, a incubadora ajuda o aluno a desenvolver técnicas de negócios e marketing.

CENTRO DE INOVAÇÃO – promove a qualificação de organizações e profissionais.

LABELO – encarregado de certificar os produtos que estão sendo desenvolvidos.

TECNOPUC – parque científico e tecnológico da PUCRS.

AGÊNCIA DE GESTÃO TECNOLÓGICA – cuida de toda a gestão do projeto e negociação com as empresas, incluindo a prestação de contas.

ESCRITÓRIO DE TRANSFERÊNCIA DE TECNOLOGIA – discute e define a propriedade intelectual de cada um dos envolvidos no projeto (aluno, empresa, professor).

AGÊNCIA DE GESTÃO DE EMPREENDIMENTOS – ajuda o pesquisador a transformar sua pesquisa em um produto para o mercado.

NÚCLEO DE APOIO À GESTÃO DA INOVAÇÃO – elabora diagnósticos de inovação.

Todas essas unidades trabalham juntas para que as parcerias com diferentes empresas possam ser proveitosas tanto para os investidores quanto para a universidade e seus alunos, sem deixar de pensar na sustentabilidade.

Avelino finalizou sua palestra destacando algumas das importantes lições aprendidas desde a criação da Rede InovaPUC:

• Ambiente favorável – O ambiente favorável gira em torno do entendimento dos processos por parte de todos os envolvidos, o que comporta desde o entendimento da universidade como um todo (englobando reitor, pró-reitores e professores) até o entendimento por parte das empresas, que tem que saber exatamente o que a universidade pode ou não fazer.

• Recursos – Em princípio não há falta de recursos, mas sim de projetos bem elaborados.
• Confiança – Tem que haver confiança para trabalhar com grandes empresas. Cuidado especial e envolvimento de profissionais especializados durante a negociação e a elaboração dos contratos.

• Gerenciamento – Sem um bom gerenciamento não é possível realizar os projetos. Sem cronograma, acompanhamento e planejamento, fica muito difícil conversar com as empresas.

• Envolvimento de ambas as partes – É muito importante estabelecer um processo de INTERAÇÃO universidade/empresa, na qual ambas são compreensivas em relação às suas limitações, dificuldades, etc.

• Conhecimento do papel de cada parte – O que a universidade e a empresa podem e não podem fazer.

• Resultados – Podem ser apresentados em forma de artigos, protótipos ou formação e capacitação.

Para encerrar, o professor Avelino respondeu algumas perguntas da plateia, a respeito da entrada de novas empresas no Parque Tecnológico, dos profissionais envolvidos em cada uma das unidades da Rede InovaPUC e sobre a forma como a PUCRS lidava com a propriedade intelectual.

9. PAINEL EMPRESAS E OS DESAFIOS DA COMPUTAÇÃO

O Painel Empresas e os Desafios da Comunicação foi formado por Karin Breitman, *General Manager* do Centro de Pesquisa e Desenvolvimento da EMC, Guilherme Wilson, Gerente de Planejamento e Controle da Fetranspor, Parahuari Branco, diretor de pesquisa, desenvolvimento e inovação, da empresa Positivo Informática, e Gabriela Gouveia Guedes Loureiro Ruberg, Chefe do Escritório de Governança da Informação do Departamento de Tecnologia da Informação do Banco Central.

9.1 Apresentação Karin Breitman

Karin Breitman abriu o painel falando sobre a EMC e o trabalho desenvolvido pela empresa. Após apresentar um panorama sobre Big Data e seu atual mercado, Karin destacou alguns dos maiores desafios enfrentados pela EMC atualmente:

Desafio Cultural – a empresa sabe mais sobre o funcionamento de determinados aparelhos do que o engenheiro que o projetou, porque conseguem identificar e antecipar possíveis falhas ou erros nos sistemas.

Dados – Karin apontou que uma das maiores dificuldades dentro da área de Big Data é encontrar profissionais qualificados para lidar com o gerenciamento dos dados, principalmente porque dentro de uma formação universitária é difícil oferecer disciplinas capazes de abranger todos os tipos de dados aos quais se pode ter acesso. Além disso, nem todos dados com os quais se poderia trabalhar estão necessariamente disponíveis para realização de pesquisas.

9.2 Apresentação Guilherme Wilson

Em seguida, Guilherme Wilson, representante da Fetranspor, falou sobre as ações de inovação no transporte público do Brasil e a importância da computação neste setor. Guilherme destaca os Sistemas de Informação ao usuário como umas das inovações mais importantes dentro do tema da computação. O primeiro Sistema desenvolvido pela Fetranspor, em parceria com a COPPE/UFRJ, foi o Vá de Ônibus[13], em 2005. A plataforma, no entanto, exigia um esforço de manutenção e uma força de trabalho que dificultava o seu mantimento.

Em 2010, a empresa detectou a necessidade de evoluir para um sistema mais simples, via web, que facilitasse o gerenciamento e sua constante atualização. Nasceu assim a INFOBUS, ferramenta conectada ao Vá de Ônibus que faz toda a gestão da base de dados de todo o Estado do Rio de Janeiro. A INFOBUS gera arquivos semanais de todo os Estado que fornece dados para outras plataformas, como, por exemplo, Google Maps e Moovit. O próximo passo seria a informação em tempo real, ou seja, informação de posicionamento dos ônibus.

Guilherme abordou ainda as evoluções tecnológicas vividas pelo setor de transportes nos últimos 5 anos, destacando o serviço do BRT[14]. Segundo Guilherme, a implantação do BRT demandou um grande investimento em tecnologia: Tecnologias aplicadas, sistemas embarcados, monitoramento por câmeras, desenvolvimento de softwares que podem prever ações e gerar alertas de segurança, sistemas integrados de informações em tempo real (equipamentos On Board que permitem a comunicação da Operação com o Centro de Controle).

A Fetranspor ainda pretende evoluir para um sistema que seja capaz de fornecer aos usuários informações de tempo de chegada dos ônibus, no entanto, ainda não foi possível desenvolver um algoritmo capaz de levar em conta o trânsito para realizar previsões mais precisas.

13http://www.vadeonibus.com.br/
14BRT vem da sigla em inglês que significa Transporte Rápido por Ônibus, e são transportes articulados que trafegam em corredores exclusivos. http://www.brtrio.com/

Por fim, Guilherme destacou que o principal desafio enfrentado pela empresa é a comunicação eficaz com o cliente com o intuito de garantir o uso dos transportes coletivos ao invés do transporte individual.

9.3 Apresentação Parahuari Branco

O próximo palestrante foi Parahuari, da Positivo Informática, e sua fala focou na pesquisa na área de educação básica desenvolvida pela empresa. Os principais desafios mencionados por Parahuari já em sua fala inicial foram:

1) Coleta dados entre as escolas de ensino fundamental no Brasil

2) Análise dos dados para transformar em ações concretas

Segundo os dados de pesquisas apresentados por Parahuari, o acesso à Internet em escolas urbanas já chega a 58%, no entanto, 48% deste acesso é feito por banda larga, o que dificulta enormemente a coleta de dados. Nesse sentido, é preciso garantir que uma Internet de qualidade chegue até as escolas, pois, sem isso, não é viável entregar conteúdo de alta qualidade aos professores e alunos e nem realizar pesquisas de Big Data. É preciso encontrar uma forma de facilitar que as indústrias que trabalham com desenvolvimento educacional possam ter um padrão para trabalhar dentro das escolas.

Parahuari citou ainda a questão da Acessibilidade na educação, mencionando a utilização de diferentes dispositivos como smartphones e tablets para auxiliar pessoas com deficiências visuais ou auditivas, e a Distorção idade/série muito presente em escolas públicas. Como o professor pode lidar melhor com esses tipos de realidades?

Por fim, foi apresentada uma pesquisa mostrando que algumas ações simples na área de computação que podem contribuir para uma revolução na educação. São elas:

1) Definir metas e ter claro o que se quer alcançar – formação continuada do professor e fornecimento de informações para realização do planejamento

2) Acompanhar de perto e continuamente o aprendizado dos alunos

3) Usar dados sobre o aprendizado para embasar ações pedagógicas

DESAFIO:

Base nacional comum x diversidade local – O professor precisa dar conta tanto do currículo nacional quanto dos aspectos regionais. Não podem se constituir em dois blocos distintos e devem perpassar transversalmente a proposta curricular. Como ajudar o professor a fazer isso?

SUGESTÕES:

O professor carece de um assistente para conseguir fazer uma coleta intensiva de dados e de uma capacitação específica para utilização de recursos tecnológicos.

O professor gasta muito tempo com a correção de provas e acaba não dando um retorno mais desenvolvido para o aluno.

É preciso haver uma troca maior de experiências. Sistema de identificação de professores que estejam se sobressaindo e possam ajudar outros professores que não estão tendo os mesmos resultados com seus alunos.

Importante criar uma base de dados sobre os alunos que permita sugestões de planejamento – para alunos com determinadas características é sugerido um determinado tipo de abordagem.

9.4 Apresentação Gabriela Ruberg

A última palestrante a realizar sua apresentação foi Gabriela Ruberg, do Banco Central. Gabriela apresentou uma síntese dos desafios enfrentados em sua área, o Escritório de Governança da Informação do Departamento de Tecnologia da Informação, principalmente no que diz respeito à divulgação de informação internamente para a sociedade.

Após uma breve apresentação do Banco Central, Gabriela enumerou os principais desafios enfrentados pela organização no trabalho com a grande quantidade de dados pelos quais são responsáveis e apontou as medidas adotadas para lidar com eles.

DESAFIOS ATUAIS:

No momento em que a informação entra no Banco, eles têm a obrigação de entender o que aquilo significa.

1) Integração de informações e processos, interna e externamente

O Banco possui cerca de 40 departamentos e lida com diversos órgãos nacionais e internacionais. Como receber essas informações com um custo mais baixo? As informações demoram a chegar e para serem processadas.

2) Volumes crescentes de informação em cenários cada vez mais dinâmicos

Como categorizar?

Como traduzir os dados?

Como se antecipar aos problemas?

Necessidade de desenvolvimento de ferramentas analíticas – Tendo em vista determinado cenário, como eu deveria agir para atingir meus objetivos?

3) Necessidade de colaboração

Não há uma linguagem de consulta de dados universal

4) Gestão de dados mestres

Curadoria da informação – Barramento de conformidade: bloco de informações que estão ok e que permitem a realização de cruzamentos. Levar em consideração a diversidade do parque tecnológico e o histórico guardado em relação a essas informações.

5) Qualidade de dados

MEDIDAS ADOTADAS:

a) Governança da informação e racionalização de processos

Programa OtimizaBC – uma hora a informação deixa de ser válida. Ela deve ser desativada? Descartada? O que fazer?

b) Ampliação do uso da plataforma de Bussiness Inteligence e de recursos para colaboração

Self-service BI

c) Desburocratização no registro das empresas

REDESIM – Lei que diz que o cidadão deve prestar informação uma vez só. Órgãos integrados.

Perguntas da Plateia

• Como vocês lidam com a segurança de informações?

- Karin: Usar Big Data para identificar os infratores. O importante é impedir a evasão, e não bloquear a entrada porque isso já se provou impossível.

- Gabriela: Não é uma questão simples, mas temos uma equipe especializada lidando com isso. A parte mais difícil é classificar, as leis não são tão simples. Temos um conjunto de práticas que tenta lidar com isso.

• Um dos maiores desafios do Seminário foi contar com a presença das empresas. O que está faltando para que a Academia possa melhor se aproximar das empresas, para que as empresas se interessem em participar, o que poderia ser um bom approach?

- Parahuari – É preciso conhecer. Nós trabalhamos muito com Lei de Informática, então estou sempre procurando formas de investir. É uma questão de conhecer os trabalhos e ter um acesso a eles. O convite já é um caminho, falta um maior diálogo.

- Gabriela – O Banco Central tem uma história de bastante colaboração com a academia. Ter um banco de contatos institucionais, num ponto de vista em médio prazo. Para viabilizar a participação é importante a antecedência.

- Guilherme – A Academia tende a aguardar que as empresas vão até ela. É importante criar uma lista de contatos.

- Karin – Todos os nossos programas de pesquisa têm relação com a universidade. Temos que desmistificar a questão dos tempos. Os tempos na universidade são diferentes. Precisa haver um modelo, no qual a indústria e a universidade sejam parceiras. Novos modelos de interação precisam ser montados e divulgados.

• Os dados da FETRANSPOR são abertos?

- Guilherme – Existe um esforço muito grande para manter esses dados armazenados. Temos fornecido dados para parceiros. Queremos entender o porquê da solicitação de uso informação, como ela vai chegar ao usuário, se ela será utilizada de modo perverso. Mas há uma possibilidade de abertura da informação, se o dado for utilizado para o bem das pessoas.

• Como fica a propriedade intelectual?

- Gabriela – Há algumas diretrizes da casa, mas tomamos muito cuidado com as informações que vamos usar.

- Parahuari – Depende do aporte que a empresa está dando ao projeto. Nas experiências com universidades públicas é algo muito complicado, pois não existe flexibilidade.

- Karin – Não temos uma regra geral, depende muito de cada caso. Em geral a EMC preserva a propriedade do parceiro. Quem fica com a propriedade intelectual deve ser alguém que vá atuar sobre ela.

10. NETWORKING DE PROJETOS

No Networking de Projetos, cada apresentador tinha 5 minutos para expor seu trabalho. Três trabalhos foram apresentados aos participantes do evento:

1. Um Ambiente para Análise dos Dados do ENADE. Autor: Reinaldo Viana – Centro Universitário Augusto Motta
2. Sistemas de computação dedicados adaptativos, distribuídos, heterogêneos e inteligentes no setor de energia. Autor: Carlos Augusto Martins, da PUC-MG
3. Questões urbanas da sociedade civil como desafios para cidades inteligentes com tecnologias sociais. Autor: André Gomes e Paulo Motta, da UNISUAM

10.1 Um Ambiente para Análise dos Dados do ENADE

O projeto teria sido originado de um trabalho de conclusão de curso dos alunos de Ciência da Computação do Centro Universitário Augusto Motta. A maior parte dos cursos de graduação do Brasil é avaliada em um ciclo trienal pelo ENADE, exame organizado pelo MEC e pelo INEP. O ENADE envolve uma participação do governo federal, das instituições de ensino superior e também envolve informações que são disponibilizadas ao cidadão. Os resultados são disponibilizados (dados abertos) e um dos objetivos da própria organização da prova seria a construção de uma série historiada capaz de permitir um diagnóstico da educação. No ENADE de 2012 469 mil estudantes, de 1625 instituições, realizaram a prova. Com posse desses dados os pesquisadores buscaram criar uma maneira mais atrativa para que as instituições e a própria sociedade pudesse ter acesso a essas informações com uma solução de BI. Foi feita uma análise que gerou um aplicativo, atualmente já funcional, que permite tirar conclusões por região do Brasil a respeito do tipo de curso avaliado, quantidade de instituições participantes, média do desempenho da prova de conhecimentos gerais, quantidade de cursos avaliados, além de permitir a comparação entre instituições e possibilitar uma série histórica de desempenho do curso.

10.2 Sistemas de computação dedicados adaptativos, distribuídos, heterogêneos e inteligentes no setor de energia

Autor: Carlos Augusto Martins, da PUC-MG

A motivação para apresentação da proposta teria sido o primeiro documento produzido sobre os grandes desafios tanto no Brasil quanto no Reino Unido e nos Estados Unidos. Esta análise permite identificar que a tendência dos Grandes Desafios nos últimos anos é o aumento da produção de dados através de dispositivos inteligentes, que, posteriormente passam a ser interligados em rede, e que, por fim, evoluem para cidades inteligentes. Como fica a educação em relação à computação e à programação? Educação fundamentada em PROJETOS, que aumenta a chance de o aluno lidar com problemas reais de formas mais simples. Baseado nos desafios identificados pela SBC em 2014, resolveram focar sua pesquisa na questão da energia e perceberam que existiriam características desejadas desses dispositivos inteligentes. É preciso que eles sejam adaptativos para que possam lidar com situações não conhecidas em tempos de projeto.

10.3 Questões urbanas da sociedade civil como desafios para cidades inteligentes com tecnologias sociais

Autor: André Gomes e Paulo Motta, da UNISUAM

Cidade Inteligente = Cidade que produz equidade – Cidade que trate a todos com suas particularidades

O que se espera de uma cidade inteligente?

- Criatividade
- Cosmopolismo
- Diversidade
- Participação na tomada de decisões
- Acesso aos serviços públicos
- Transparência
- Coesão social

Questões que também deve ser consideradas quando se pensa nas cidades inteligentes do futuro.

Exemplos de investimentos em tecnologia social:

❖ Sistemas de dados abertos: LAI

❖ Prontuário Nacional

❖ Sistemas ubíquos de recomendação de serviços públicos

❖ Sistemas ubíquos de recomendação de mapeamento de serviços

Investimento para empoderar a sociedade sobre a administração pública. O sistema tem que se adequar ao cosmopolismo e à diversidade da cidade, e não o contrário.

11. ANEXOS/ARTIGOS PREMIADOS

CIENCIA DE DADOS

Fábio Porto e Artur Ziviani

Resumo. Discutimos ciência de dados como um desafio da computação para os próximos anos. Esta proposta está relacionada primordialmente ao desafio "Gestão da Informação em grandes volumes de dados multimídia distribuídos" com aplicação amplamente interdisciplinar, cobrindo múltiplos domínios do eixo ciência-indústria-governo transversalmente.

INTRODUÇAO

O tratamento do dilúvio de dados sendo produzido pelas ciências e por bilhões de usuários de serviços de Internet globais se apresenta como um dos grandes desafios para a atual sociedade do conhecimento [Bell et al., 2009]. Apresentado de forma geral como um vetor de múltiplas facetas, o fenômeno ainda está sendo interpretado pelos cientistas e vem impulsionando iniciativas em diversas áreas. Nas ciências, o dilúvio apareceu como a expressão de uma nova maneira de investigação [Wright, 2014], incitando biólogos, astrônomos, físicos, e demais pesquisadores em diversas áreas científicas, a enfrentarem problemas computacionais na chamada e-ciência, que se tornam barreiras para as suas descobertas. Na indústria, o dilúvio de dados aparece fortemente como análise preditiva [Dhar, 2013] em sintonia com o ambiente de computação em nuvem, provendo escalabilidade e tolerância a falhas, em ambientes computacionais cada vez mais complexos e de tamanho proporcional ao desafio. Na setor governamental, há oportunidades de se debruçar sobre imensas bases de dados do setor público com vistas a gerar planejamento mais eficiente bem como novos serviços que possam melhorar o atendimento ao cidadão. Novas profissões especializadas no trato e, principalmente, na análise e interpretação de grandes volumes de dados, surgiram, trazendo o método científico para o setor empresarial.

Neste contexto, constitui-se um desafio técnico-científico em computação o estudo metódico para a extração generalizada e em escala de conhecimento relevante a partir de uma imensa massa de dados, em geral dinâmicos [Jagadish et al., 2014]. A abordagem

Laboratório Nacional de Computação Científica (LNCC) Av. Getúlio Vargas, 333 – 25651-075 – Petrópolis, RJ

{fporto,ziviani}@lncc.br

a esse desafio com aplicações em diversas áreas no eixo ciência-indústria-governo emerge como uma nova espécie de ciência. A chamada *Ciência de Dados* incorpora elementos variados e se baseia em técnicas e teorias oriundas de muitos campos básicos em engenharia e ciências básicas, sendo assim intimamente ligada com muitas das disciplinas tradicionais bem estabelecidas, porém viabilizando uma nova área altamente interdisciplinar. Dessa forma, associado a este espírito de aplicação interdisciplinar, a ciência de dados emerge como componente cada vez mais importante nas mais diversas áreas, tais como saúde, petróleo, energia, financeira, esporte, astronomia, bioinformática, Internet, mobilidade urbana, defesa cibernética, comunicação móvel e biodiversidade, apenas para mencionar algumas.

Em ambiente altamente interdisciplinar com aplicações em áreas tão distintas, emerge o *grande desafio* comum às aplicações nessas tão diversas áreas de se identificar os princípios, métodos e técnicas fundamentais para o gerenciamento e análise de grandes volumes de dados, suplantando as dificuldades inerentes ao grande volume de dados em análise [Jacobs, 2009, Lazer et al., 2014]. Especificamente, identificamos três linhas de pesquisa principais cujo amadurecimento acreditamos deverá conduzir rumo à consolidação da área de ciência de dados em um horizonte de alguns anos de pesquisa e desenvolvimento: (i) gerência de dados; (ii) análise de dados; e (iii) análise de redes complexas; todos essas linhas considerando a larga-escala dos dados a serem analisados bem como seu dinamismo. A partir da pesquisa básica nesses aspectos fundamentais de análise de dados em larga-escala, há também um grande potencial tecnológico na pesquisa aplicada em ciência de dados com impacto em diferentes áreas do conhecimento e de setores de atuação ao longo do eixo ciência-indústria-governo.

Um desafio correlato se torna a formação de recursos humanos altamente qualificados no desenvolvimento de pesquisa básica e aplicada na fronteira do conhecimento em ciência de dados. Esse *cientista de dados* possui demanda crescente no eixo ciência -indústria-governo [Davenport e Patil, 2012]. Esse profissional tem uma expectativa de formação tipicamente sólida em ciência da computação e aplicações, modelagem, estatística, analítica e matemática, além do conhecimento mínimo do domínio de aplicação.

Em suma, enfrentar de forma fundamental o grande desafio da ciência de dados permite contribuir a melhor posicionar o Brasil na direção da nova ciência baseada em dados, preparando recursos humanos altamente qualificados, e desenvolvendo o alicerce para sua projeção de forma relevante na sociedade do conhecimento.

O restante deste documento está organizado como segue. Na Seção 2 apresentamos a motivação desta proposta de grande desafio em ciência de dados, tendo por base diferentes cenários de aplicação. Delineamos os desafios de pesquisa básica em ciência de dados nas três principais linhas de pesquisa identificadas ao longo da Seção 3.

Os desafios relacionados à formação de recursos humanos na área de ciência de dados são apresentados na Seção 4. Finalmente, a Seção 5 traz algumas considerações finais.

MOTIVAÇAO PARA PESQUISA BASICA E APLICADA EM CIÊNCIA DE DADOS

A grande motivação para nossa proposta de desafio relacionado à ciência de dados emerge da experiência anterior em realizar atividades de pesquisa e desenvolvimento em gestão e análise de dados, bem como análise de redes complexas, em cenários de aplicação das áreas mais diversas. Exemplos são astronomia, biodiversidade, Internet, petróleo & gás, saúde e comunicação móvel. Nesta seção, apresentamos uma descrição sucinta da relevância de ciência de dados nesses cenários de aplicação, já em investigação pelos autores.

Essa experiência anterior, portanto, permitiu a identificação de um clamor por pesquisa básica nos aspectos fundamentais de análise de dados em larga-escala, o principal ponto de motivação para a nossa proposta de ciência de dados como grande desafio à computação nos próximos anos. Isso também traz consigo a qualificação e justificativa da relevância deste desafio dado o espectro amplo de impacto e aplicação de avanços de ciência de dados ao longo das linhas de pesquisa delineadas nestes cenários de aplicação, bem como em outros oriundos do eixo ciência-indústria-governo.

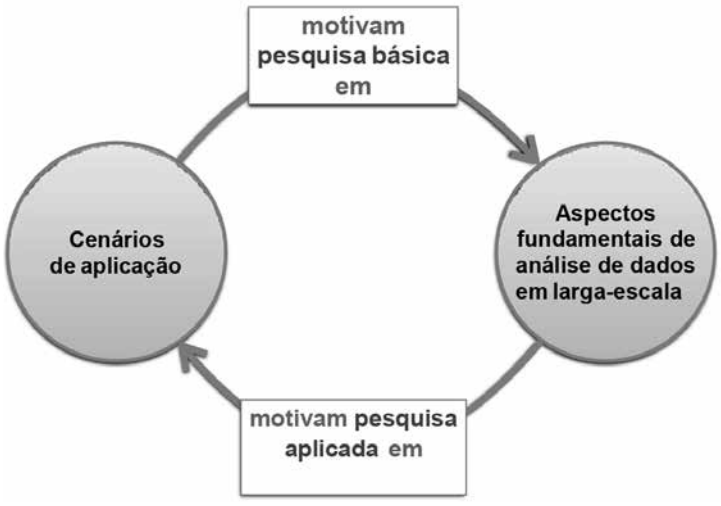


Figura 1. Motivação cíclica para pesquisa básica e aplicada em ciência de dados.

As diferentes aplicações práticas de ciência de dados, tais como os cenários de aplicação ilustrativos descritos nesta seção, ao mesmo tempo em que são alvos para elaboração de novas soluções em pesquisa aplicada, muitas vezes propiciam a oportunidade de elaboração de novos arcabouços teóricos em pesquisa básica, de caráter mais geral, para a solução dos problemas práticos. A Figura 1 ilustra esse ciclo de motivação para a pesquisa básica e aplicada em ciência de dados. Essa abordagem que liga teoria e prática é uma das estratégias gerais adotadas pelos autores, que tem as suas pesquisas centradas na análise de dados em diferentes campos. Exemplos ilustrativos de cenários de aplicação atuais de ciência de dados em áreas diversas, nos quais os autores tem experiência, são:

Astronomia: O LNCC é membro do Laboratório Inter-institucional de eAstronomia (LIneA),[1] onde tem-se gerenciado e processado dados obtidos de grandes levantamentos astronômicos. Estes levantamentos produzem dados a partir de imagens telescópicas fotografadas por instrumentos terrestres. A partir das imagens, corpos celestes são identificados e suas características anotadas produzindo um conjunto de dados chamado Catálogo Astronômico. Tais catálogos podem abrigar até centenas de bilhões de objetos celestes. Processar tal volume incomum de dados desses catálogos de forma eficiente requer seu particionamento e alocação distribuída em um cluster. Estratégias de particionamento devem atender a requisitos de diferentes dataflows de análise, dificultando a determinação de critérios adequados a distintas aplicações. Encontrar estratégias que coincidam com os critérios de dataflows é um problema em aberto. A integração de diversos catálogos, produzidos por diferentes levantamentos, também traz o problema de resolução de entidades, uma vez que a identificação de objetos estrelares é feita com base em sua posição, cuja medida varia em diferentes telescópios [Freire et al., 2014].

Biodiversidade: Para monitorar as mudanças na biodiversidade é essencial coletar, documentar, armazenar e analisar indicadores sobre a distribuição espaço temporal das espécies, além de obter informações sobre como elas interagem entre si e com o ambiente em que vivem [Michener et al., 2012]. Nesse contexto, o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr)[2] visa integrar e disseminar dados coletados e publicados por diversas instituições brasileiras. O SiBBr, cuja infraestrutura computacional está sediada no LNCC, desempenha também o papel de nó brasileiro do Global Biodiversity Information Facility (GBIF).[3] O SiBBr já permite a agregação de dados de espécies e ocorrências disponibilizadas por diversas instituições acadêmicas e de pesquisa bem como de órgãos governamentais. Um primeiro protótipo de workflow científico para modelagem de distribuição de espécies [Gadelha Jr. et al.,

1http://www.linea.gov.br

2012b] permite uma execução escalável e com registro de informações de proveniência [Gadelha Jr. et al., 2012a].

Internet: A Internet apresenta grandes desafios para a caracterização de sua estrutura e comportamento [Chen, 2001]. De fato, a Internet mostra-se atualmente como um conjunto de redes complexas interdependentes entre si, abrangendo desde as redes de comunicação que formam a infraestrutura básica de interconexão até redes sociais online envolvendo bilhões de usuários, passando por redes no nível aplicativo de troca de conteúdo. Há, portanto, grandes desafios para a caracterização, análise e modelagem de tais redes na Internet, bem como a rede WWW sobre esta, pois esses estudos devem também preservar a privacidade de usuários, o que impõe desafios adicionais à coleta eficiente e detalhada de informações importantes para condução de pesquisa. Os autores contam com experiências diversas em diferentes aspectos da coleta, análise e modelagem do imenso volume de dados envolvidos na investigação da estrutura e comportamento das atuais redes complexas que são formadas na e pela Internet bem como o impacto em suas aplicações [Gueye et al., 2006, Freire et al., 2008, Ziviani et al., 2012, Las-Casas et al., 2013].

Petróleo e Gás: A pesquisa de petróleo e gás em áreas profundas é um grande desafio no Brasil, com grandes campos em profundidades de mais de 5 kms na área do pré-sal. A investigação nestes campos envolve a captura de reflexos de ondas sísmicas enviadas a partir da superfície. Ondas enviadas em direção às camadas subaquáticas são refletidas por camadas rochosas no fundo do mar e recapturadas por sensores na superfície. Uma vez capturadas e processadas para limpeza dos dados, os chamados traços sísmicos são combinados em um grande conjunto de dados representando a região investigada. A atividade de analisar os sinais sísmicos para detecção de feições de interesse é chamada de interpretação geofísica e tem valor econômico bastante relevante. Neste sentido, o desenvolvimento de técnicas que possam apoiar a detecção de falhas sobre campos muito grandes, como o pré-sal brasileiro, é um problema cujas soluções estão ainda em sua infância. Além do problema básico da gerência de grande volume de dados, a inferência de feições a partir de sinais em ondas sísmicas é um grande desafio.

Saúde: A área de saúde lida rotineiramente com enormes quantidades de dados. Esse volume de dados somente aumenta devido à adoção crescente de sistemas de informação em saúde e prontuários eletrônicos do paciente. O LNCC tem experiência na área de sistemas de informação em saúde [Correa et al., 2011, Gomes et al., 2012], além da instituição ser a atual sede do INCT-MACC (INCT em Medicina Assistida por Computação Científica).[4] Há grandes desafios na gestão e análise de dados ligados à

[2]http://www.sibbr.gov.br 3http://www.gbif.org

área de saúde, tais como a agregação, manutenção, interoperabilidade, interpretação desses dados, sem mencionar questões de privacidade devido à evidente sensibilidade dos dados [Nambiar et al., 2013]. A tendência é de ainda maior expansão no volume de dados num futuro próximo devido ao uso crescente de sensores ou mesmo dispositivos móveis para coleta de dados individualizados em ambientes residenciais ou pré-hospitalares [Estrin, 2014]. Outra tendência recente é a abordagem de modelagem por redes complexas de problemas relativos à área de saúde, seja relacionando doenças [Barabási et al., 2011], seja relacionando serviços de saúde para melhor coordenação de cuidados e uso de recursos de forma centrada no paciente [Pretz, 2014].

Comunicação móvel: Dados coletados de redes de telefonia celular tem um enorme potencial de prover informações valiosas sobre o relacionamento dinâmico de indivíduos [Eagle et al., 2009] ou sobre mobilidade humana [Becker et al., 2013] a um custo relativamente baixo e numa escala sem precedentes. A análise de enormes volumes de dados de redes celulares hoje apresenta impacto em diversas áreas, de melhor planejamento e dimensionamento das próprias redes de telecomunicação até mais indiretamente, por exemplo, planejamento urbano [Iqbal et al., 2014]. O LNCC tem experiência, com colaboradores, no estudo de dados de redes celulares para a investigação da dinâmica da carga da rede e mobilidade humana devido a eventos de larga-escala em ambiente urbano [Xavier et al., 2012, Xavier et al., 2013]. São desafios nesse cenário de aplicação a gestão e análise dos dados em grande volume, assim como a análise das redes complexas de relacionamento que emergem tipicamente desse tipo de dado.

DESAFIOS DE PESQUISA EM CIÊNCIA DE DADOS

Esta proposta de desafio em ciência de dados situa-se diretamente relacionada ao tema estratégico de tecnologias da informação e comunicação (TICs), tendo relação imediata com o primeiro dos programas prioritários para os setores portadores de futuro, tal como definido na Estratégia Nacional de Ciência, Tecnologia e Inovação (ENCTI) 2012-2015,[5] publicada pelo Ministério da Ciência, Tecnologia e Inovação (MCTI). De fato, gestão de informação em grandes volumes de dados já era reconhecido em 2006 com um dos grandes desafios para a computação brasileira para os 10 anos seguintes no primeiro relatório a respeito realizado pela Sociedade Brasileira de Computação (SBC). Esses desafios são hoje ainda maiores vistos os atuais volumes de dados a analisar, bem como seu dinamismo e capilaridade, que surgem como oportunidades de progresso científico e inovação tecnológica em diferentes áreas do eixo ciência-indústria-governo.

[4] http://macc.lncc.br
[5] http://www.mct.gov.br/upd_blob/0218/218981.pdf

Figura 2. Desafios de pesquisa básica em ciência de dados para diversos cenários de aplicação no eixo ciência-indústria-governo.

Nesta seção, as linhas de pesquisa identificadas como portadoras de desafios em ciência de dados para os próximos anos são detalhadas, sendo essas: (i) gestão de dados;

(ii) análise de dados; e (iii) análise de redes complexas; todos essas linhas considerando a larga-escala dos dados a serem analisados bem como seu dinamismo. Avaliadas de modo integrado, essas linhas de pesquisa encontram-se nos desafios apresentados ao se confrontar com os grandes volumes de dados produzidos atualmente nas ciências, governo e indústria. A Figura 2 ilustra essa visão de ciência de dados como central a diversas áreas do eixo ciência-indústria-governo, tendo as linhas de pesquisa propostas como ponto de ligação entre essas áreas e ciência de dados. Acreditamos não serem essas linhas de pesquisa absolutamente exaustivas quanto aos desafios de ciência de dados. Porém estas cobrem em grande parte os aspectos fundamentais da pesquisa básica e aplicada em ciência de dados em larga-escala de forma transversal a diversos cenários de aplicação no eixo ciência-indústria-governo, o foco principal motivador de nossa proposta de desafio.

GESTAO DE DADOS

Representacão de dados

O sucesso de Sistemas de Bancos de Dados Relacionais aplicados a problemas comerciais reduziu a representação de dados a tabelas bidimensionais. Em se tratando de gerência de grandes volumes de dados, a propriedade de sua representação se reflete positiva-

mente no desempenho do acesso aos dados e, por sua vez, espelha necessidades de aplicações mais complexas do que aquelas apoiadas por bancos de dados relacionais. Assim, domínios que têm sido em grande parte negligenciados pelo suporte de banco de dados, tais como simulações numéricas, análises sísmicas e redes de interação gênica, para citar apenas alguns de muitos domínios nesta área, requerem representações de dados em sintonia com representações complexas, tais como: espaço, tempo, grafos e seqüências.

Sistemas como o SciDB,[6] por exemplo, propõem a representação de dados científicos em matrizes multidimensionais, o que de fato se apresenta como uma estrutura interessante, na medida em que generaliza o conceito de sistemas de coordenadas como índice de valores de variáveis calculadas. Vimos, no entanto, em [Costa et al., 2012], que o modelo apresenta problemas na representação de malhas irregulares e que mais esforços precisam ser dedicados nessa área. Igualmente, dados do tipo sísmico do tipo *pos-stack* apresentam como estrutura uma lista de seqüências, representando traços sísmicos, e ordenadas no espaço 2D e 3D. Efetuar consultas em dados deste tipo é um grande desafio e pode alavancar pesquisas importantes na área de petróleo, por exemplo. Ainda nesta linha, dados de acompanhamento de atletas [Porto et al., 2012] podem ser modelados por trajetórias virtuais, onde cada ponto reflete o valor de um elemento observável de interesse em múltiplas dimensões, incluindo o estado de treinamento, o atleta sendo observado e a data da observação. O mesmo modelo [Spaccapietra et al., 2008] tem sido usado em soluções de mobilidade urbana para representação de indivíduos em regiões urbanas, a partir de dados de mobilidade obtidos durante o uso de telefones celulares.

Um outro modelo que tem sido bastante utilizado em contextos bastante distintos como redes sociais e biologia de sistemas utiliza a representação em grafos, formando uma rede complexa (ver Seção 3.3). Interações humanas ou as produzidas pela expressão de certos genes podem ser fielmente retratadas em associações entre nós de um grafo, onde estes espelham os objetos do domínio de interesse (i.e. pessoas ou genes) e aqueles referem-se aos tipos de interações ocorrendo entre esses objetos. Características dos objetos ou das interações podem ser retratadas em suas propriedades, de forma semelhante aos atributos de relações. Na medida em que o conjunto de dados que se pretende representar em grafos aumenta, problemas como o desempenho de consultas apontam para estratégias semelhantes às adotadas em sistemas relacionais, como o particionamento. No entanto, a natureza de consultas em grafos baseada na navegação através das associações tornam a definição de critérios de particionamento de dados mais difícil.

/www.scidb.org

Finalmente, de forma mais comum, uma mesma aplicação pode requerer a representação de parte de seus dados em diferentes modelos. Assim, retornando ao cenário de biologia de sistemas, uma aplicação de inferência de redes de interação gênica pode utilizar o modelo de grafos para representar as interações, o modelo relacional para características dos genes e organismos envolvidos e um modelo XML para armazenamento de dados de proveniência do processo de inferência.

Um desafio é a pesquisa de estruturas para representação de dados, como acima discutidas, tendo por exemplo as aplicações discutidas na Seção 2 como cenários motivadores da investigação. O objetivo principal seria identificar as lacunas na adoção de tais representações, principalmente tendo como alvos a expressividade para aplicações e o alcançado desempenho para grandes volumes de dados.

Tratamento de incerteza em dados

A crescente disponibilização de informações, seja na web ou capturadas por instrumentos e sensores, potencialmente subsidia tomada de decisões mais precisas. Isto ocorre em parte pelo maior conhecimento do fenômeno observado, seus estados limiares e casos espúrios. No entanto, conforme discutido em [Srivastava, 2012], informações contraditórias disponibilizadas de forma independente produzem o resultado oposto, confundindo os usuários. Neste aspecto, graus de confiança podem ser atrelados às fontes de informação, caracterizando-as e permitindo que sejam selecionadas a partir de um modelo de custo. No contexto de dados científicos, a acurácia da informação pode depender do instrumento sendo utilizado para sua captura. Em cenários de múltiplos instrumentos, semelhantes a diferentes fontes de dados, a discrepância entre caracterizações de objetos comuns a várias fontes dificulta uma possível integração e torna complexa respostas a consultas. Finalmente, em modelos de simulação computacional, a incerteza própria dos modelos e dos parâmetros precisa ser identificada e informada ao usuário para tomada de decisão [Gonçalves e Porto, 2014].

Em geral, tais exemplos mostram uma faceta de dados em grandes volumes associados à imprecisão da informação. Enquanto em situações mais controladas os dados se mostram bem comportados, em contextos em que dados são capturados de diversas fontes autônomas, ou que advêm de processos por natureza imprecisos tais como modelos computacionais, é preciso equipá-los com características de incerteza e propagar essa representação nas inferências deles extraídas [Suciu et al., 2011].

Neste tópico, um desafio é se avaliar o tratamento de incerteza em representações como as discutidas na Seção 3.1.1. Em modelos de matrizes multi-dimensionais, por exemplo, a incerteza pode se distribuir ao longo do espaço-tempo e novos mecanismos de inferência probabilística podem ser necessários. Igualmente, no cálculo de inferência

de redes de interações, as associações se estabelecem de forma imprecisa e será interessante avaliar a atribuição de incerteza em nós e arestas de grafos.

Particionamento de dados

O processamento de grandes volumes de dados, para que se torne escalável, requer o particionamento de dados entre nós de um cluster de computadores. O problema de particionamento de dados é antigo [Özsu e Valduriez, 2011], no entanto os tipos de aplicações que acessam grandes volumes de dados é de natureza fundamentalmente distinta das aplicações convencionais: (i) aplicações varrem grandes volumes de dados; (ii) estratégias de acesso variam devido à característica exploratória do processo de pesquisa; (iii) dados não sofrem atualizações. Neste contexto, as estratégias de particionamento precisam ser reavaliadas. No conhecido framework Hadoop,[7] pedaços de tamanho uniforme, sem conotação semântica, estabelecem as fronteiras de cada unidade de alocação (i.e., partição). Fica evidente que estratégias mais próximas das características dos dados devem favorecer tanto o armazenamento quanto o acesso.

Neste sentido, [Curino et al., 2010] propõem uma estratégia de particionamento chamada Schism, que se baseia na análise de grafos em que nós representam os objetos e arestas o acesso conjunto daqueles. Dessa forma, algoritmos como *min-cut* encontram partições que minimizam o acesso em mais de uma partição (i.e. maximizam o acesso local). Apesar de interessante, as estratégias derivadas de Schism não são adequadas para o particionamento de petabytes de dados e, muito menos, para acesso de varredura de boa parte dos dados, como é o caso freqüente em grandes volumes de dados. Por outro lado, o processamento em dataflow requer mudança na estratégia de particionamento, uma vez que deve atender a uma seqüencia de atividades do dataflow. A combinação de paralelismo de tarefas em dataflows com particionamento de dados nas fontes, e aqueles produzidos por etapas intermediárias do dataflow, sugere que novas estratégias para o particionamento de dados devam ser avaliadas.

Neste tópico de pesquisa, há o desafio de se investigar estratégias para o armazenamento de grandes volumes de dados, tendo como foco técnicas para particionamento, replicação e indexação de dados.

Processamento de grandes volumes de dados

A necessidade de se utilizar programas ad-hoc para o processamento intensivo de dados norteou o desenvolvimento de modelos de processo baseados em dataflows, cujo maior expoente é o paradigma MapReduce [Dean e Ghemawat, 2008] e sua implementação

[7] http://hadoop.apache.org

aberta Hadoop. Processos em dataflows se diferem do processamento de consultas tradicionais em vários aspectos: semântica de transformação de dados desconhecida; programas e dados fora do alcance dos SGBDs; otimização reduzida; estatísticas escassas, apenas para comentar alguns desses aspectos. Ainda assim, alguns trabalhos tentam recuperar para dataflows os benefícios de otimização automática e gerência de proveniência [Ogasawara et al., 2011, Hueske et al., 2012]. Neste contexto, a coincidência entre os critério de interesse de dados expresso pelas atividades e aquele referente ao do particionamento de dados devem dirigir a estratégia de execução. Modelos de execução norteados a processos intensivos de CPU, tal como o modelo de ativação, podem ser integrados ao dirigido a dados formando um modelo de execução heterogêneo apropriado à diferentes etapas do dataflow.

Neste tópico de pesquisa, é um desafio investigar estratégias e algoritmos para o processamento eficiente de grandes volumes de dados por dataflows, assim como novos sistemas tipo NoSQL [Mohan, 2013], privilegiando aspectos específicos de cada aplicação e dos dados nela armazenados.

Análise de dados

Processo de análise de dados

Em linhas gerais, a análise de dados corresponde a um conjunto de atividades que devem ser desempenhadas, desde a seleção dos dados até a produção do conhecimento, que é o principal produto da análise. A análise de dados envolve o processamento de coleções de objetos em busca de padrões consistentes, de forma a detectar relacionamentos sistemáticos entre variáveis componentes desses objetos e gerar conhecimento não facilmente detectado. Dá-se o nome de processo de análise de dados à especificação do encadeamento desse conjunto de atividades. As atividades que compõem o processo de análise de dados podem ser organizadas em quatro etapas: seleção, pré-processamento, métodos de análise e avaliação [Han et al., 2006].

O processo de análise pode ser compreendido como um caso particular de experimentação científica *in-silico* [Stevens et al., 2007], no qual os dados são volumosos, as estruturas de dados precisam ser bem definidas e os métodos de análise de dados são computacionalmente intensivos. Neste contexto, é apropiado estabelecer um tratamento datacêntrico a esses experimentos, compreendendo a desafios diretamente ligados aos apresentados na Seção 3.1.1. A pesquisa envolve, então, estrutura de dados e algoritmos para apoiar tanto às etapas (seleção de dados, pré-processamento, algoritmos de mineração e análise), quanto ao processo de análise de dados como um todo.

No que tange ao processo de análise de dados, há também uma necessidade preemente de utilizar processamento de alto desempenho (PAD) para se conseguir realizar a

análise de dados em larga-escala. Além dos desafios mencionados na Seção 3.1.4 sobre o processamento desses grandes volumes de dados, há outros importantes desafios no estabelecimento desses processos. Esses processos são comumente modelados como workflows [Goderis et al., 2006]. Nestes workflows, as atividades e dados estão direcionados a execução em algum ambiente de PAD (clusters, computação em nuvem) [Oliveira et al., 2010, Ogasawara et al., 2013], onde ocorre a decomposição destes workflows em dataflows e, tem-se a alocação destes dataflows e seus respectivos dados aos recursos computacionais.

Em função da diversidade de plataformas existentes, um dos grandes desafios é estabelecer uma representação deste workflow que seja agnóstica ao meio em que será executado e, ao mesmo tempo, possibilite a otimização de sua execução no ambiente alvo, considerando-se os aspectos mencionados nas Seções 3.1.3 e 3.1.4. Diante deste cenário repleto de desafios para a execução de workflows de análise em larga-escala, tem-se oportunidades para explorar técnicas e métodos de acesso para grande volumes de dados, as quais possam lidar, principalmente, com os problemas relacionados com o particionamento, distribuição, movimentação e sumarização de dados presentes nos experimentos. Novamente, essas técnicas e métodos dependem da plataforma na qual os dados são processados.

Um desafio neste tópico é investigar métodos para lidar com replicas parciais do banco de dados, ao mesmo tempo em que desenvolveremos métodos que tirem vantagem das infraestruturas virtualizadas de processamento em larga-escala para reduzir o tempo de acesso aos dados processados e persistidos em memória principal.

Técnicas de análise em grandes volumes de dados

A análise de dados propriamente dita é apoiada por um conjunto de métodos que incluem tanto os métodos tradicionais de mineração de dados — pré-processamento, classificação, predição, agrupamento, associação e visualização — quanto os métodos de modelagem computacional [Liao et al., 2012]. Esses métodos, por sua vez, são apoiados por técnicas clássicas, dentre as quais incluem-se k-means, Partitioning Around Medoids (PAM), árvores de decisão, redes neurais, Support Vector Machines (SVM) e Apriori.

A partir dessas técnicas de análise clássicas, diversos algoritmos, implementações, adaptações e variações estão presentes em ferramentas de análise de dados consolidadas, como, por exemplo, a linguagem R. O desafio consiste tanto na correta aplicação desses algoritmos, como também na implementação adequada para se atingir a escalabilidade desses algoritmos nos processamento de grandes volumes de dados. Essas aplicações devem ser encapsuladas em wrappers para fazer parte dos experimentos de análise (workflows). Nesse contexto, um aspecto muito importante para

a ciência de dados, consiste em como preparar os dados para a aplicação destas técnicas. A correta aplicação das técnicas de normalização, transformação [Ogasawara et al., 2010], remoção de outliers [Gupta et al., 2014], seleção de atributos e definição de amostras, pode significar a diferença entre obter ou não conhecimento e produzir valor agregado. Outro aspecto fundamental consiste em se ter uma infraestrutura que possibilite explorar as diferentes técnicas para selecionar a mais adequada para os dados trabalhados. No processo de condução da ciência de dados, isso consiste em visualizar os resultados parciais e poder ajustar parâmetros nas técnicas de análise durante a execução do experimento [Mattoso et al., 2013].

Análises orientadas a hipóteses

A análise de dados em larga-escala, como advogado nessa proposta de desafio, requer um processo científico no qual o cientista de dados formula hipóteses e utiliza os dados disponíveis, ou desenvolve um experimento *in-silico* para sua produção, como base para validação. Neste contexto, permitir que hipóteses científicas sejam gerenciadas computacionalmente amplia o suporte computacional ao ciclo-de-vida da ciência *in-silico*. De fato, o novo cientista de dados exige dos sistemas computacionais o mesmo aparato encontrado pela ciência tradicional em seus laboratórios. Por um lado, dados observacionais são capturado[5] de forma digital e formam a base para formulação de hipóteses sobre os fenômenos sendo investigados. Por outro lado, sistemas de workflows fornecem o ferramental para levar a cabo a validação das hipóteses seja através de simulações computacionais, cujos resultados são confrontados com as observações, seja na determinação de padrões nos dados que estatisticamente estabeleçam relações causais propostas pela hipótese científica.

Neste sentido, vários trabalhos têm sido propostos para apoio na criação e validação de hipóteses de forma automática [Schmidt e Lipson, 2009]. A ferramenta Eureqa[8] obteve grande sucesso, produzindo através de um algoritmo de programação dinâmica expressões matemáticas que modelam um conjunto de dados fornecido, sem de fato apresentar semântica física correspondente.

No contexto de pesquisas orientadas à formulação de hipóteses que reflitam a interpretação do fenômeno observado, o problema é abordado de maneira top-down. Um modelo teórico inicial é proposto, possivelmente como um conjunto de equações matemáticas. Utilizando-se de métodos numéricos, deriva-se sua representação equivalente em forma computacional. A avaliação do modelo computacional produz os dados a partir dos quais as hipóteses podem ser avaliadas. Em [Haas, 2014], um sistema de banco de dados para apoio ao cálculo de simulações numéricas é apresentado. O sistema permite que atualizações e progressos na análise de hipóteses sejam realizados inteiramente baseadas em dados. A partir de sua geração, dados de simulação

computacional passam a estar disponíveis para sua avaliação analítica. Em [Gonçalves e Porto, 2014], estes são carregados em um sistema de bancos de dados probabilísticos onde a característica de incerteza dos modelos associados é inferida. O cálculo da incerteza associada aos dados de simulação permite a adoção de modelos Bayesianos, que através do cálculo de probabilidade condicional permite integrarmos dados observacionais com dados simulados, a priori.

A ciência de dados orientada a hipóteses pode tomar, igualmente, uma abordagem menos teórica. Em sistemas de recomendação, por exemplo, nos quais ferramentas de busca na web fazem sugestões automaticamente a seus usuários, encontra-se um modelo ad-hoc relevante de formulações de hipóteses e de sua avaliação sobre a grande quantidade de dados disponíveis. Neste contexto, pode-se avaliar e valorar milhões de hipóteses simultaneamente, ordenando-as e escolhendo a que melhor reflita a natureza dos dados.

A expressão e validação de hipóteses científicas *in-silico* está na base da realização da ciência em dados. Seja através de expressões formais matemáticas ou de determinação de padrões em dados, ela é fundamental no avanço desta nova disciplina. A compreensão e adoção do métodos baseados em hipóteses *in-silico* está, no entanto, na sua infância. A compreensão das diferentes facetas de sua representação e de modelos mais ou menos formais influencia o processo científico e precisa ser melhor compreendido. Do ponto de vista do suporte computacional, modelos, técnicas e algoritmos devem ser concebidos para abrigar a ciência em dados vista desta forma. Neste contexto, há um claro desafio relacionado a se aprofundar a compreensão sobre essas diversas facetas do processo científico *in-silico*, ampliando o suporte computacional à ciência praticada em dados.

Análise de redes complexas

Habitamos um mundo extremamente conectado e esta conectividade em permanente crescimento impacta nossas vidas de maneiras que ainda não compreendemos totalmente [Vespignani, 2009]. Simultaneamente, novas ferramentas, métodos e tecnologias nos permitem atualmente um potencial sem precedentes de extrair conhecimento de enormes volumes de dados de alguma maneira interconectados em diversos campos da ciência, de redes sociais a redes biológicas.

No campo bastante ativo de redes complexas, mais recentemente também chamado de *ciência de redes* [Kocarev e In, 2010], o desenvolvimento de um grande conjunto de atividades de pesquisa nos últimos 10-15 anos foi muito estimulado pela disponibilidade

[8]http://www.nutonian.com/products/eureqa

crescente de dados empíricos e o aumento correspondente na capacidade computacional para analisar tais dados. Isso permitiu a percepção de similaridades nas estruturas de redes oriundas de áreas bastante distintas, o desenvolvimento de uma série de ferramentas e métodos para caracterizar e modelar tais redes, bem como o entendimento do impacto da estrutura dessas redes nos processos dinâmicos que ocorrem nessas redes. Esse desenvolvimento acelerado da disponibilidade de dados e aplicações imediatas com base nesses colabora para a atual demanda por pesquisa básica nos aspectos fundamentais de análise de redes complexas.

Nesse contexto, diferentes sistemas de grande porte, tanto naturais quanto artificiais, com elementos diversos interconectados podem ser representados por meio de redes complexas de larga-escala [Albert e Barabási, 2002, Newman, 2003] (ver Seção 3.1.1, onde discute-se a possibilidade de representação por grafos). A adoção dessa abordagem baseada em redes complexas para modelagem atualmente impacta, não somente áreas científicas, mas também diversos setores do governo ou da indústria como base para aplicações estratégicas em definição de políticas públicas ou sistemas de recomendação, respectivamente, apenas para nomear algumas aplicações. Podemos identificar duas frentes relacionadas com a caracterização, análise e modelagem de redes complexas, uma é lidar com a dimensão que as mesmas tipicamente tem adquirido em diversas áreas e outra frente é lidar com o dinamismo das redes complexas atuais, sendo que este último pode estar associado a redes complexas dinâmicas com larga-escala. Discutimos essas frentes a seguir.

Redes complexas de larga-escala

A caracterização,análise e modelagem de redes complexas de larga-escala [Albert e Barabási, 2002, Rosvall e Bergstrom, 2010] é um desafio chave no domínio de ciência de dados, dada a vasta presença maciça e escala das redes complexas com as quais várias áreas do conhecimento lidam atualmente. Um dos desafios nesse tema é investigar algoritmos, métodos ou técnicas que permitam a análise de características globais de redes complexas, muitas vezes muito custosas computacionalmente ou mesmo intratáveis dependendo da escala, por métricas locais, viabilizando a análise dessas redes, mesmo que aproximadamente [Everett e Borgatti, 2005, Wehmuth e Ziviani, 2013]. Esse contexto se aplica a áreas científicas diversas, bem como sistemas naturais ou artificiais modelados por redes complexas presentes em desafios na ciência, indústria e também no setor governamental. Lidar com redes complexas de larga-escala, mesmo estáticas, impõe desafios em todas as subáreas de pesquisa relacionadas à gestão e análise de dados (em interseção com as outras duas linhas de pesquisa consideradas nesta proposta e discutidas nas Seções 3.1 e 3.2).

Dinamismo em redes complexas de larga-escala

O principal desafio na linha de pesquisa de análise de redes complexas reside no estabelecimento de fundações sólidas para a caracterização, análise e modelagem de redes complexas dinâmicas de larga-escala. Redes complexas dinâmicas podem apresentar um dinamismo espaço-temporal. Esse dinamismo pode ser variante no tempo (i.e., arestas e nós variam ao longo do tempo, podendo ser representados por grafos variantes no tempo [Holme e Saramäki, 2012, Wehmuth et al., 2014b]; variante no espaço, onde múltiplas redes interdependentes podem ser associadas em camadas (podendo ser representadas por redes multicamadas [Kurant e Thiran, 2006, De Domenico et al., 2013]; ou mesmo ambos [Kivelä et al., 2014, Wehmuth et al., 2014a].

Dado esse dinamismo das redes complexas que emergem em diferentes cenários de aplicação, tais como os descritos na Seção 2, esse é um desafio chave para o avanço de aspectos fundamentais na área de ciência de dados ao lidar com sistemas naturais ou artificiais modelados por redes complexas de larga-escala, sobretudo dinâmicas. Esse desafio se projeta tanto na caracterização, análise e modelagem da dinâmica da estrutura dessas redes complexas, mas também na caracterização, análise e modelagem dos processos dinâmicos que ocorram sobre essas redes complexas. A análise se torna ainda mais desafiadora em cenários combinados onde se requer a análise de processos dinâmicos em execução sobre estruturas de redes dinâmicas, requerendo portanto a caracterização, análise de modelagem *de* e *em* redes complexas dinâmicas de larga -escala. Exemplos de processos dinâmicos são a difusão de informação, identificação de comunidades de interesse, partição dos grafos, ou detecção de anomalias.

Associado à investigação dos aspectos fundamentais de análise do dinamismo de e em redes complexas de larga-escala, também torna-se um desafio a pesquisa aplicada para o desenvolvimento de métodos, técnicas e ferramentas que sejam de relevância prática em cenários de aplicação reais, tais como os discutidos na Seção 2.

FORMAÇAO DE RECURSOS HUMANOS

A formação de recursos humanos é um dos grandes desafios mais importantes para o avanço da área de ciência de dados no Brasil. A pesquisa básica e aplicada envolve tipicamente o desenvolvimento de técnicas, metodologias, modelos, algoritmos e arquiteturas em ciência de dados. O perfil profissional de cientista de dados possui demanda crescente no eixo ciência-indústria-governo [Davenport e Patil, 2012]. Esse profissional tem uma expectativa de formação tipicamente sólida em ciência da computação e aplicações, modelagem, estatística, analítica e matemática, além do conhecimento mínimo do domínio de aplicação, dada sua atuação intrinsecamente interdisciplinar. Assim, o novo profissional em ciência de dados reúne um conjunto de competências interdisciplinares dificilmente encontradas no profissional formado pelos cursos verti-

cais atualmente oferecidos nas universidades. No Brasil, começam a aparecer alguns poucos cursos curtos de especialização, mas sua estruturação em cursos lato-sensu, de graduação e pós-graduação ainda é incipiente, se tanto.

Há, portanto, o grande desafio de formação de recursos humanos altamente qualificados em pesquisa básica e aplicada na fronteira do conhecimento em ciência de dados.

CONSIDERAÇOES FINAIS

O avanço tecnológico das últimas décadas culminou com a capacidade de obtenção e geração de imensos volumes de dados, tanto de fenômenos naturais quanto de sistemas artificiais. O novo cenário delineado nesse contexto abre em realidade novas necessidades, perspectivas e oportunidades de avanços tecnológicos relacionados ao desenvolvimento de técnicas, metodologias, modelos, algoritmos e arquiteturas para se fazer frente ao desafio de analisar e interpretar esses imensos volumes de dados que emergem em aplicações de diversas áreas do conhecimento. Há, portanto, um grande potencial tecnológico na pesquisa básica e aplicada em ciência de dados, tal como aqui discutido, dado o foco nos aspectos fundamentais da análise de dados em larga-escala com impacto em diferentes áreas de conhecimento básico bem como cenários de aplicação.

Ciência de dados é uma área recente tanto no Brasil quanto no exterior. Há, entretanto, já algumas iniciativas recentes em instituições de ponta no exterior que focam em ciência de dados. Alguns exemplos são o Data Science Institute[9] no Imperial College, o Institute for Data Sciences & Engineering[10] na Columbia University, o Berkeley Institute for Data Science (BIDS)[11] na UC Berkeley, o Center for Data Science (CDS)[12] da New York University ou a iniciativa de recrutamento expressivo em Data Science a partir de junho/2014 na Boston University.[13] É, portanto, relevante concentrar esforços para enfrentar o grande desafio de ciência de dados em nosso país, contribuindo para preencher a lacuna existente atualmente no Brasil nesta área.

Em suma, é necessário posicionar o Brasil na direção da nova ciência baseada em dados, enfrentando os desafios de pesquisa básica e aplicada em ciência de dados, preparando recursos humanos altamente qualificados na área, de forma a desenvolver o alicerce para a projeção do país de forma relevante e em bases sólidas na sociedade do conhecimento.

[9]http://www3.imperial.ac.uk/data-science
[10]http://idse.columbia.edu
[11]http://vcresearch.berkeley.edu/datascience 12http://cds.nyu.edu
[13]http://www.bu.edu/provost/2014/06/23/university-provosts-faculty-hiring-initiative-in-data-science

AGRADECIMENTOS

Os autores agradecem FAPERJ, CNPq, FINEP e MCTI pelo apoio. Os autores agradecem a Eduardo Ogasawara (CEFET-RJ) por sua contribuição na parte de análise de dados.

REFERENCIAS

Albert, R. e Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Barabási, A.-L., Gulbahce, N., e Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., e Volinsky, C. (2013). Human mobility characteriza- tion from cellular network data. *Communications of the ACM*, 56(1):74–82.

Bell, G., Hey, T., e Szalay, A. (2009). Beyond the Data Deluge. *Science*, 323:1298–1298.

Chen, T. M. (2001). Increasing the observability of internet behavior. *Communications of the ACM*, 44(1):93–98.

Correa, B. S., Gonçalves, B., Teixeira, I. M., Gomes, A. T., e Ziviani, A. (2011). Atoms: a ubiquitous teleconsultation system for supporting ami patients with prehospital thrombolysis. *International journal of telemedicine and applications*, 2011:2.

Costa, R. G., Porto, F., e Schulze, B. (2012). Towards analytical data management for numerical simulations. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, pages 210–214.

Curino, C., Jones, E., Zhang, Y., e Madden, S. (2010). Schism: a workload-driven approach to database replication and partitioning. *Proceedings of the VLDB Endowment*, 3(1-2):48–57.

Davenport, T. H. e Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century.

Harvard Business Review.

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., e Arenas, A. (2013). Mathematical formulation of multilayer networks. *Physical Review X*, 3(4):041022.

Dean, J. e Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64– 73.

Eagle, N., Pentland, A. S., e Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278.

Estrin, D. (2014). small data, where n = me. *Communications of the ACM*, 57(4):32–34. Everett, M. e Borgatti, S. P. (2005). Ego network betweenness. *Social networks*, 27(1):31-38.

Freire, E. P., Ziviani, A., e Salles, R. M. (2008). Detecting voip calls hidden in web traffic. *Network and Service Management, IEEE Transactions on*, 5(4):204–214.

Freire, V. P., Macedo, J. A. F., e Porto, F. (2014). NACluster: A non-supervised clustering algorithm for matching multi catalogs. In *The e-Science Workshop for Work In Progress, IEEE International Conference on e-Science*.

Gadelha Jr., L. M., Wilde, M., Mattoso, M., e Foster, I. (2012a). MTCProv: a practical provenance query framework for many-task scientific computing. *Distributed and Parallel Databases*, 30(5-6):351–370.

Gadelha Jr., L. M. R., Stanzani, S., Correa, P., Dalcin, E., Gomes, C. R. O., Sato, L., e Siqueira, M. (2012b). Scalable and provenance—enabled scientific workflows for predicting distribution of species. In *Proc. 8th International Conference on Ecological Informatics (ISEI 2012)*, Brasília, DF.

Goderis, A., Li, P., e Goble, C. (2006). Workflow discovery: the problem, a case study from e-science and a graph-based solution. In *International Conference on Web Services (ICWS)*, pages 312–319. IEEE.

Gomes, A. T. A., Ziviani, A., Correa, B. S. P. M., Teixeira, I. M., e Moreira, V. M. (2012). SPLiCE: a software product line for healthcare. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 721–726. ACM.

Gonçalves, B. e Porto, F. (2014). -DB: Managing scientific hypotheses as uncertain data. In *Proc. of the Very Large Data Bases (VLDB)*.

Gueye, B., Ziviani, A., Crovella, M., e Fdida, S. (2006). Constraint-based geolocation of internet hosts. *Networking, IEEE/ACM Transactions on*, 14(6):1219–1232.

Gupta, M., Gao, J., Aggarwal, C., e Han, J. (2014). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267.

Haas, P. J. (2014). Model-data ecosystems: challenges, tools, and trends. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 76–87. ACM.

Han, J., Kamber, M., e Pei, J. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Holme, P. e Saramäki, J. (2012). Temporal networks. *Physics reports*, 519(3):97–125. Hueske, F., Peters, M., Sax, M. J., Rheinländer, A., Bergmann, R., Krettek, A., e Tzoumas, K. (2012). Opening the black boxes in data flow optimization. *Proceedings of the VLDB Endowment*, 5(11):1256–1267.

Iqbal, M. S., Choudhury, C. F., Wang, P., e González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52(8):36. Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., e Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., e Porter, M. A. (2014).

Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Kocarev, L. e In, V. (2010). Network science: A new paradigm shift. *IEEE Network*, 24(6):6–9.

Kurant, M. e Thiran, P. (2006). Layered complex networks. *Physical review letters*, 96(13):138701.

Las-Casas, P. H., Guedes, D., Almeida, J. M., Ziviani, A., e Marques-Neto, H. T. (2013). Spades: Detecting spammers at the source network. *Computer Networks*, 57(2):526–539.

Lazer, D., Kennedy, R., King, G., e Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, 343(6176):1203–5.

Liao, S.-H., Chu, P.-H., e Hsiao, P.-Y. (2012). Data mining techniques and applications—a decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311.

Mattoso, M., Ocaña, K., Horta, F., Dias, J., Ogasawara, E., Silva, V., de Oliveira, D., Costa, F., e Araújo, I. (2013). User-steering of HPC workflows: State-of-the-art and future directions. In *Proceedings of the 2nd ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, page 4. ACM.

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Kel- ling, S., Koskela, R., Tenopir, C., e Vieglais, D. A. (2012). Participatory design of DataO-

NE—enabling cyberinfrastructure for the biological and environmental scien- ces. *Ecological Informatics*, 11:5–15.

Mohan, C. (2013). History repeats itself: sensible and nonsensql aspects of the nosql hoopla. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 11–16. ACM.

Nambiar, R., Bhardwaj, R., Sethi, A., e Vargheese, R. (2013). A look at challenges and opportunities of big data analytics in healthcare. In *IEEE International Conference on Big Data*, pages 17–22.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P., e Mattoso, M. (2011). An algebraic approach for data-centric scientific workflows. *Proceedings of the VLDB Endowment*, 4(12):1328–1339.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., Oliveira, D., Porto, F., Valduriez, P., e Mattoso, M. (2013). Chiron: a parallel engine for algebraic scientific workflows. *Concurrency and Computation: Practice and Experience*, 25(16):2327–2341.

Ogasawara, E., Martinez, L. C., de Oliveira, D., Zimbrão, G., Pappa, G. L., e Mattoso, M. (2010). Adaptive normalization: A novel data normalization approach for non-stationary time series. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Oliveira, D., Ogasawara, E., Baião, F., e Mattoso, M. (2010). Scicumulus: A lightweight cloud middleware to explore many task computing paradigm in scientific workflows. In *IEEE International Conference on Cloud Computing*, pages 378–385. IEEE.

Özsu, M. T. e Valduriez, P. (2011). *Principles of distributed database systems*. Springer.

Porto, F., Moura, A. M., Silva, F. C., Bassini, A., Palazzi, D. C., Poltosi, M., Castro, L. E. V., e Cameron, L. (2012). A metaphoric trajectory data warehouse for olympic athlete follow-up. *Concurrency and Computation: Practice and Experience*, 24(13):1497–1512.

Pretz, K. (2014). Better health care through data. Tech Focus, The Institute, IEEE. Rosvall, M. e Bergstrom, C. T. (2010). Mapping change in large networks. *PloS one*, 5(1):e8694.

Schmidt, M. e Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., e Vangenot, C. (2008). A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146.

Srivastava, D. (2012). Towards analytical data management for numerical simulations. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*.

Stevens, R., Zhao, J., e Goble, C. (2007). Using provenance to manage knowledge of in silico experiments. *Briefings in bioinformatics*, 8(3):183–194.

Suciu, D., Olteanu, D., Ré, C., e Koch, C. (2011). Probabilistic databases. *Synthesis Lectures on Data Management*, 3(2):1–180.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425.

Wehmuth, K., Fleury, E., e Ziviani, A. (2014a). On multiaspect graphs. *arXiv preprint arXiv:1408.0943*.

Wehmuth, K. e Ziviani, A. (2013). DACCER: Distributed assessment of the closeness centrality ranking in complex networks. *Computer Networks*, 57(13):2536–2548.

Wehmuth, K., Ziviani, A., e Fleury, E. (2014b). A Unifying Model for Representing Time-Varying Graphs. Technical Report RR-8466, INRIA.

Wright, A. (2014). Big data meets big science. *Communications of the ACM*, 57(7):13–15.

Xavier, F. H. Z., Silveira, L., Almeida, J., Malab, C., Ziviani, A., e Marques-Neto, H. T. (2013). Understanding human mobility due to large-scale events. In *3rd Conference on the Analysis of Mobile Phone Datasets (NetMob)*.

Xavier, F. H. Z., Silveira, L. M., Almeida, J. M. d., Ziviani, A., Malab, C. H. S., e Marques-Neto, H. T. (2012). Analyzing the workload dynamics of a mobile phone network in large scale events. In *Proceedings of the first workshop on Urban networ- king (URBANE), ACM CoNEXT*, pages 37–42. ACM.

Ziviani, A., Cardozo, T. B., e Gomes, A. T. A. (2012). Rapid prototyping of active measurement tools. *Computer Networks*, 56(2):870–883.

SISTEMA DE INFORMAÇÃO EM SAÚDE SILVESTRE – "SISS-GEO"

Marcia Chame1, Helio J. C. Barbosa2, Luiz Gadelha2
Douglas A. Augusto1, Eduardo Krempser2, Livia Abdalla1

1. SOLUÇÕES COMPUTACIONAIS PARA OS DESAFIOS DA MODELAGEM DE EMERGÊNCIA DE DOENÇAS ORIUNDAS DA FAUNA SILVESTRE

As alterações ambientais, incluindo as mudanças climáticas e a perda da biodiversidade, são fatores determinantes para a emergência de doenças oriundas de animais silvestres [6] e podem estar na origem das forças seletivas de novas variações genéticas que permitem o rompimento de barreiras biológicas por agentes patogênicos e o aumento do potencial de dispersão de doenças em humanos. Embora não consideradas adequadamente nas políticas de vigilância em saúde, o quadro é relevante, uma vez que a maioria (60,3%) das doenças infecciosas circula entre animais e humanos (zoonoses), das quais 71,8% dessas são causadas por patógenos com origem na vida silvestre [12].

Essas emergências quase sempre estão associadas aos territórios mais atingidos por impactos naturais e antropogênicos, compondo também a gama de parâmetros que tornam as desigualdades sociais ainda mais severas e injustas, como forte repercussão e custos para a saúde e a qualidade de vida (UNEP/CDB/SBSTTA/18/17)15. Nos últimos 15 anos diversos estudos mostraram o efeito de diluição da biodiversidade na dispersão de agentes patogênicos e na modulação de sua dinâmica de sua transmissão [13, 26, 18]. No entanto, os estudos e ações no último século, apesar da expansão do conhecimento epidemiológico, reagiram a eventos de emergência de doenças específicas na população humana, com algumas tentativas de mitigação. Considerando a baixa capacidade de reverter as mudanças climáticas e os impactos ambientais determinados pelo crescimento humano, além de nossa forma de produ-

[1]Fundação Oswaldo Cruz
Programa Institucional Biodiversidade de Saúde (Fiocruz – PIBS)
Rio de Janeiro – RJ – Brasil
[2]Laboratório Nacional de Computação Científica (LNCC/MCTI)
Petrópolis – RJ – Brasil
mchame@fiocruz.br, hcbm@lncc.br, lgadelha@lncc.br
daa@fiocruz.br, krempser@lncc.br, abdallalivia@fiocruz.br

[15]http://www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-17-en.pdf

ção e consumo de recursos naturais, parece razoável prever que não conseguiremos deter a emergência destas doenças. Esse quadro é paradoxal em países megadiversos, como o Brasil. Ao mesmo tempo em que a riqueza de espécies existentes traz a elas associadas também a riqueza de parasitos e, portanto, um potencial risco, é esta complexidade de espécies e de suas relações que protegem e estabiliza a dinâmica das transmissões, reduzindo o surgimento de surtos de doenças, um dos mais importantes serviços ecossistêmicos. Diante deste cenário, mais que buscar respostas eficientes para situações de crise, há motivos para se buscar ações que antecipem problemas para que se possa mitigá-los quando possível, e responder rapidamente a eles quando prevenção e/ou mitigação falharem.

Essa abordagem vem sendo fortalecida com programas internacionais, como o "One world, one health" da OMS/OIE16 e o Plano Estratégico 2011-2020 da Convenção da Diversidade Biológica (CDB)17 e, estrategicamente, em programas governamentais de países desenvolvidos que já envidam recursos e esforços consideráveis para o rastreamento de patógenos em todo o mundo, quer seja para prevenção de pandemias, como as recém ocorridas com as novas gripes e Ebola, o desenvolvimento de novos fármacos ou mesmo por preocupações de guerra biológica. No Brasil são incipientes as estratégias sistematizadas para o monitoramento e a previsão de ocorrências de doenças advindas da biodiversidade, que seguem modelo de notificação de agravos já ocorridos em humanos, insuficientes para ações preventivas [2].

2. ESCOPO DA SOLUÇÃO PROPOSTA

As relações que unem a biodiversidade à saúde são complexas porque frequentemente são indiretas, dispersas no espaço e no tempo e dependentes de inúmeras forças [19]. Não se trata somente de identificar espécies e sua distribuição geográfica. No contexto da emergência de zoonoses estão imbricadas diversas espécies de patógenos, vetores e hospedeiros que modulam evolutivamente entre si, dinâmicas e composição populacional que reagem às mudanças ambientais [13].

Enfrenta-se, portanto, um desafio de múltiplas dimensões. A primeira é sensibilizar os tomadores de decisão sobre a necessidade de monitorar a circulação de patógenos na fauna silvestre antes que estes acometam humanos, ampliando as ações da vigilância em saúde para além dos humanos. A segunda dimensão é construir mecanismo que não se limite

[16]http://www.oneworldonehealth.org/
[17]http://www.cbd.int/sp/

frente ao tamanho territorial do Brasil, às políticas setoriais pouco integradas, às urgências nacionais que absorvem pessoal e não se ocupariam das tarefas do monitoramento. A terceira é como integrar múltiplas competências, já que esse mecanismo deverá abarcar especialistas para lidar com dados, espécies e contextos sociais e ambientais distintos. A quarta é como efetivamente obter informações e tratá-las adequadamente. A quinta é extrair dos dados informações relevantes e identificar realmente os riscos e prevê-los e, por fim, o compromisso de levar informações relevantes à sociedade.

Como evidenciado, a coleta de dados, o monitoramento e a extração de conhecimento e informações sobre a saúde silvestre e suas relações com a saúde humana mostram-se tarefas desafiadoras envolvendo inúmeras áreas do conhecimento, as caracterizando como atividades interdisciplinares que visam à modelagem de um sistema dinâmico e complexo.

É também evidente que grandes áreas da computação são de indispensável aplicação no contexto apresentado, tais como a modelagem computacional, a aprendizagem de máquina e a programação paralela, porém suas aplicações não são óbvias, dada à necessidade de integração de informações de diferentes meios, a complexidade e dimensionalidade dos dados a serem manipulados e a sensibilidade envolvida na utilização e divulgação desses dados.

Na construção de um sistema capaz de tratar todas as questões elencadas, elaborou-se a colaboração entre a Fundação Oswaldo Cruz (Fiocruz) e o Laboratório Nacional de Computação Científica (LNCC).

Considerando as tecnologias disponíveis e a sinergia estabelecida entre as duas instituições, o desenvolvimento do Sistema de Informação em Saúde Silvestre – SISS-Geo[18] é a proposta para avançar sobre os desafios postos. Sua concepção busca a integração e a participação de diversos segmentos da sociedade, desde o registro de dados primários por qualquer pessoa interessada, na aplicação do conceito de ciência cidadã, ao diagnóstico confiável de agentes patogênicos que circulam na fauna silvestre com potencial de acometimento humano com a participação de rede de laboratórios e especialistas, até os desafios computacionais e matemáticos que incluem sistemas analíticos e de predição; mineração de dados; processos intensivos; programação paralela; integração de sistemas, dados (desestruturados e heterogêneos) e informações; geoprocessamento; aprendizagem de máquina, meta-heurísticas e visualização de dados para a construção de modelos de alerta e previsão de agravos advindos da biodiversidade e promovidos pelas forças motrizes antropogênicas.

[18]http://www.biodiversidade.ciss.fiocruz.br/apresentacao-0
[19]http://www.biodiversidade.ciss.fiocruz.br

O SISS-Geo tem como característica essencial o tratamento dos seus dados em ambiente espacialmente referenciado. Tem como objetivos: (i) proporcionar, de maneira rápida e eficiente, o fluxo de informações entre o Centro de Informação em Saúde Silvestre19 da Fiocruz e o sistema nacional de vigilância em saúde, com contribuição especial ao Centro de Informações Estratégicas em Vigilância em Saúde – CIEVS/MS; as redes participativas em saúde silvestre e de laboratórios; a população em geral que deseja participar do processo e; os diferentes centros de monitoramento da biodiversidade, como o MCTI (Ministério da Ciência, Tecnologia e Inovação), ICMBio (Instituto Chico Mendes de Biodiversidade), JBRJ (Jardim Botânico do Rio de Janeiro), MAPA (Ministério da Agricultura, Pecuária e Abastecimento), Embrapa (Empresa Brasileira de Pesquisa Agropecuária), etc.; (ii) criar, a partir dos dados e informações georreferenciadas, modelos de alerta e previsão de agravos à saúde silvestre e humana, de modo a atuar como sistema sentinela para doenças emergentes e reemergentes e, ainda, disponibilizar os resultados das modelagens espaciais para a comunidade científica e tomadores de decisão; (iii) permitir meios adequados para integração do sistema georreferenciado com bancos de dados geográficos de parceiros governamentais e não governamentais; (iv) adequar-se ao padrão de metadados da Infraestrutura Nacional de Dados Espaciais (INDE)20, visando disponibilizar, com eficiência e total compatibilidade, dados relacionados à saúde silvestre para a comunidade científica e a população em geral.

SISS-Geo está construído sobre quatro macro-módulos. O primeiro sistematiza a captação dos registros georreferenciados das observações de campo de animais e de suas condições físicas e do ambiente ao seu redor, feita por colaboradores por meio de dispositivos móveis (Android e IOS) e em ambiente Web, os organizando em bancos de dados (Seção 3.2). O segundo gera, utilizando-se da modelagem a partir de dados, modelos automatizados de alertas considerando as distâncias territoriais, os intervalos de tempo entre elas, a similaridade dos grupos taxonômicos envolvidos, com notabilidade para primatas, quirópteros, roedores e carnívoros, mas não a eles limitados, as condições físicas de encontro dos animais no campo, de acordo com padrões clínicos pré-categorizados, além das informações ambientais do local onde o animal foi avistado (Seção 3.3.1).

A partir da indicação de importância e emergência gerada pelo modelo de alerta, busca-se a integração de atores da Rede Participativa em Saúde Silvestre e, especialmente, da Rede de Laboratórios em Saúde Silvestre, e dos serviços em saúde e ambiental instituídos no País, para a coleta de amostras biológicas em animais no campo e a confiabilidade do diagnóstico.

[20]http://www.inde.gov.br

O diagnóstico confiável realimentará e validará o modelo de alerta e, a partir da correlação inicial das condições ambientais de ocorrência, espera-se estudos e geração de modelos de previsão de oportunidades ecológicas para a ocorrência de doenças oriundas da biodiversidade, o que constitui o terceiro módulo (Seção 3.3.2).

Finalmente, o quarto módulo contempla o desafio do entendimento das relações que governam o fenômeno em questão, a partir dos modelos treinados. Neste contexto, a extração de conhecimento atua como principal mecanismo de sugestão de hipóteses para posterior investigação/validação do especialista (Seção 3.3.3).

A automatização da busca de padrões de ocorrência visa tornar possível e eficiente a abrangência de informações das pessoas mais simples até especialistas e em todo o território nacional, gerar conhecimento sobre o entendimento de padrões possíveis e parâmetros que contribuem para a ocorrência de doenças, a formação, em médio e longo prazo, de pesquisadores capazes de desenvolver modelagens complexas na área da ecologia das doenças e sua gestão integrada à tecnologia de informação geográfica, e obviamente, gerar dados para a política nacional de saúde e de conservação da biodiversidade.

A proposta inspirou-se no desejo de tornar público e buscar reforços para uma longa caminhada que congrega pesquisadores, especialistas de múltiplas áreas e a sociedade para que, por meio da computação, a informação e ações de prevenção de doenças cheguem as regiões mais remotas do País. Surge da prática de muitos anos de pesquisa de campo, no semiárido brasileiro, onde informações relevantes de agravos em animais silvestres foram perdidas ou dispersas e a falta de sua sistematização impossibilitou ações importantes tanto para a contenção de doenças em humanos, quanto para a conservação das espécies.

O SISS-Geo nasce dos esforços em criar ações inovadoras e integradas para a transversalização da biodiversidade nos setores do País. Integra-se às ações da Fundação Oswaldo Cruz no "Projeto de Ações Público-privadas para a Biodiversidade" – PROBIOII21, coordenado pelo MMA, e desenvolvido pelo FUNBIO, Embrapa, MAPA, MS, MCTI, Jardim Botânico do Rio de Janeiro, ICMBio e Fiocruz. O LNCC se juntou ao projeto da Fiocruz e garantiu sua realização numa parceria de construção de conhecimento de longo prazo.

Ações correlatas como a Rede de Laboratórios em Saúde Silvestre, que conta com 43 laboratórios nas diversas regiões do Brasil e da Rede Participativa, com mais de 1000 seguidores no Facebook e a realização da 1ª Conferência Brasileira em Saúde Silvestre e Humana [2] fundamentam o escopo da solução proposta.

[21]http://www.mma.gov.br/biodiversidade/projetos-sobre-a-biodiversidade/projeto-nacional-de-acoes-publico-privadas-para-biodiversidade-probio-ii

3. DESAFIOS E PROPOSTAS RELATIVAS AO SISS-GEO

Os principais desafios computacionais encontrados no SISS-Geo podem ser categorizados em quatro classes: gerenciamento de dados sobre saúde silvestre, geoprocessamento, aprendizagem de máquina e ferramentas de apoio à rastreabilidade e à composição de análises sobre saúde silvestre, detalhados a seguir.

3.1 Gerenciamento de Dados sobre Saúde Silvestre

Para monitorar as mudanças na biodiversidade é essencial coletar, documentar, armazenar e analisar indicadores sobre a distribuição espaço-temporal das espécies, além de obter informações sobre como elas interagem entre si e com o ambiente em que vivem [15]. O desenvolvimento e implantação de mecanismos para produzir esses indicadores dependem do acesso a dados confiáveis obtidos em expedições de campo, por sensores automáticos, em coleções biológicas e na literatura acadêmica. Esses dados normalmente estão disponíveis em diversas instituições que utilizam formatos e identificadores distintos, o que torna desafiador o trabalho de integração de dados.

As metodologias e técnicas usadas para gerenciar e analisar esses dados definem uma área de pesquisa frequentemente chamada de Informática na Biodiversidade [23, 11]. Algumas iniciativas para o estabelecimento de padrões de metadados e de publicação de dados, como o EML [7] e o Darwin Core [25], conseguiram estabelecer conjuntos de identificadores para descrever os principiais conceitos sobre biodiversidade. Embora esses identificadores cubram apenas uma fração dos conceitos possíveis, eles permitem que instituições publiquem seus dados sobre biodiversidade utilizando o mesmo formato e que estes sejam coletados e processados automaticamente por sistemas agregadores, como o GBIF22.

Por meio da utilização desses padrões, o SISS-Geo poderia coletar dados de ocorrências de espécies disponibilizados por diversos provedores, assim como oferecer os dados armazenados no seu próprio banco de dados para a comunidade em geral em um formato de fácil consumo. O Darwin Core tem sido estendido para inclusão de conceitos sobre temas específicos, como informações sobre interações e polinizadores (Interaction Extension to Darwin Core) e sobre fichas de espécies (Plinian Core). Seria importante avaliar e propor uma extensão do padrão para contemplar informações sobre saúde silvestre nos registros de observação de espécies, o que normalmente é realizado no contexto do TDWG23.

22http://www.gbif.org
23http://www.tdwg.org

3.2 Geoprocessamento

A espacialização e visualização geográfica são hoje condições básicas para a gestão da informação. Quase nunca ela é simples por questões que incluem a necessidade de normalização, atualização e acesso a dados qualificados. Nos estudos das doenças infecciosas a espacialização dos dados ainda precisa considerar pulsos e flutuações populacionais determinados por diversos fatores como sazonalidade, períodos reprodutivos, migrações, entre outros [16].

O Sistema de Informação em Saúde Silvestre – SISS-Geo tem por objetivo construir informações relevantes e confiáveis, capazes de cooperar nos processos decisórios do Ministério da Saúde, do Ministério da Agricultura, Pecuária e Abastecimento e do Ministério do Meio Ambiente, fornecendo subsídios para tomadas de decisão mais ágeis e oportunas.

Por se tratar de um projeto inovador, as tarefas desenvolvidas não são simples e não existem soluções prontas. É, portanto, necessária, a construção de novas metodologias e a utilização de diferentes tipos de tecnologias geográficas capazes de atender as expectativas e objetivos do SISS-Geo. A Infraestrutura de Geoprocessamento (IG) do SISS-Geo tem importância estratégica nesse processo, existindo a necessidade de superar desafios pertinentes ao controle da qualidade de dados geográficos, minimização dos erros posicionais dos modelos, espacialização da modelagem baseada em aprendizagem de máquina e disponibilização dos modelos sob a forma de mapas dinâmicos na Internet.

A modelagem de oportunidade ecológica de doenças do SISS-Geo irá utilizar uma densa massa de dados geográficos com escalas, sistemas de referência, fontes e metodologias de mapeamento distintas. Para isso é necessária a normalização e integração dos dados em banco de dados geográfico. Este será utilizado tanto no consumo de informações/dados, quanto no armazenamento dos resultados pertinentes à modelagem, sob a forma de modelos geograficamente distribuídos. Os dados de entrada para a modelagem são obtidos em função da sobreposição dos registros de animais silvestres com as bases de dados ambientais, sociais e de impactos antrópicos. Em função da localização dos registros, serão estabelecidos relacionamentos espaciais do tipo “está dentro”, “está próxima”, “intersecta”, etc.

As bases de dados sistemáticas disponibilizadas por fontes oficiais do governo Federal, Estadual e Municipal, são produzidas, em sua maioria, em pequena e média escala (1:1.000.000, 1:500.000). O mapeamento nestas escalas proporciona somente a visão geral do espaço, com grau de detalhamento e precisão reduzidos. Isso pode influenciar significativamente na modelagem do SISS-Geo, pois implicará em grau de incerteza entre o conjunto de pontos registrados e os dados cartográficos nacionais, à medida que estes forem relacionados.

A medida de incerteza geralmente corresponde ao Padrão de Exatidão Cartográfico (PEC), cujo valor é estimado para cada mapeamento e define a classificação de uma carta. No entanto, o uso da PEC é questionável quando se trata de cartografia digital [21], cujo desenvolvimento introduziu novas técnicas de mapeamento e de cálculos de erros. A Especificação Técnica para a Aquisição de Dados Geoespaciais Vetoriais (ET-ADGV) adotada na Infraestrutura Nacional de Dados Espaciais (INDE), também aborda essa questão e define novos parâmetros a serem seguidos em relação ao mapeamento sistemático do Brasil. Essa norma considera que a exatidão na aquisição do dado é igual a do produto cartográfico digital final, porque após a aquisição vetorial de um elemento qualquer, sua geometria não é mais alterada nos processos posteriores. Além disso, os padrões de exatidão considerados nessa norma são bem mais rigorosos que os baseados em cartografia analógica e são calculados com base na comparação estatística entre medições realizadas em campo e no produto digital. A adoção ET-ADGV é uma tendência, mas ainda está sob processo de adaptação, de modo que poucos dados terão essa informação documentada. Portanto, a referência do valor de exatidão posicional dos dados para as consultas espaciais do SISS-Geo será inicialmente baseada na PEC.

Para a minimização do efeito do erro posicional nos modelos do SISS-Geo será considerado a tolerância nos cruzamentos espaciais, com base na exatidão posicional dos dados em sobreposição, utilizando como referência a PEC. Busca-se com isso, estabelecer modelos com qualidade posicional suficiente para apoiar as tomadas de decisão oriundas de políticas de saúde pública.

A infraestrututra de geoprocessamento necessita também disponibilizar, ao domínio público24, os resultados e os modelos de alerta e previsão do SISS-Geo, excetuando-se as informações sensíveis25. Portanto, segue em desenvolvimento a adequação do sistema de informação geográfica para ambiente Web, que irá disponibilizar os resultados do SISS-Geo sob a forma de mapas dinâmicos/interativos e estatísticas gráficas na Internet. Como vantagem dessa tecnologia, encontra-se a facilidade de manipulação, análise e interpretação dos modelos pelo usuário final; independência de sistema operacional e interação com sistemas desktop ou outros sistemas da Internet (interoperabilidade).

3.3 Aprendizagem de Máquina

3.3.1 Agrupamento de registros de observações e predição de alerta

Quando a observação de um animal silvestre, sua condição física e ambiente circundante é registrada no SISS-Geo, seja por especialistas ou colaboradores do sistema,

[24]Lei de acesso à informação: http://cidadao.mpf.mp.br/acesso-a-informacao
[25]Instrução Normativa da INDA: http://dados.gov.br/instrucao-normativa-da-inda

esta é reunida com outros registros relacionados (previamente comunicadas) dando origem a coleção de acontecimentos que caracterizam um fenômeno. Esta é a fase de agrupamento e, embora possa soar descomplicada, envolve o desafio de se conceber/treinar modelos dotados de capacidade discriminativa em reconhecer similaridades e dissimilaridades entre eventos, baseando-se em critérios como distância espacial e temporal entre os registros, similaridade entre espécies e nas condições físicas reportadas e outros. Este fluxo de aprendizagem é sintetizado na **Figura 1**.

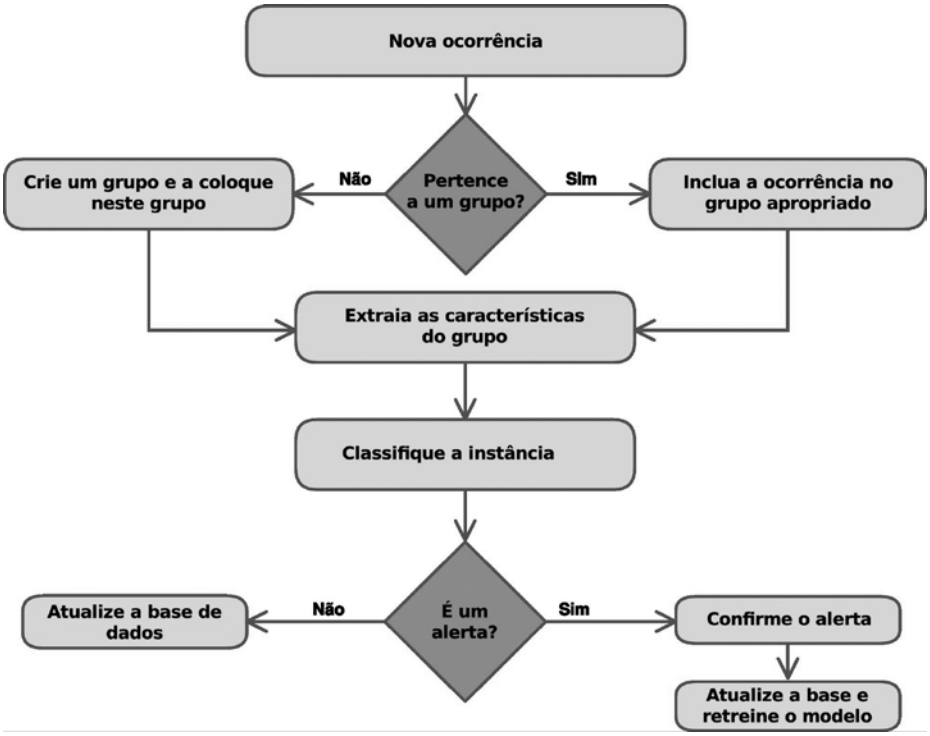


Figura 1: Fluxo relativo à aprendizagem de máquina do SISS-Geo.

A segunda parte consiste em modelar características dos registros das observações que as tornam menos ou mais relevantes, isto é, treinar o modelo de alertas. Significa predizer a gravidade dos registros de acordo com as informações trazidas pelos eventos bem como o contexto geográfico/ambiental. Por exemplo, um registro envolvendo isoladamente um animal com sintomas é em geral menos grave do que ocorrências contendo eventos similares, mas abrangendo grupos de animais. Naturalmente, em situações reais a caracterização de uma situação de alerta costuma ser bem menos óbvia, usualmente considerando vários fatores para a tomada de decisão.

Percebe-se que as atividades acima mencionadas referem-se à tarefa de agrupamento e classificação de dados, típicas da aprendizagem de máquina, e notoriamente conhecidas pela ampla pluralidade de abordagens e metodologias. São assim tarefas complexas, tanto pela natureza como também pelo grande volume de dados esperado para o sistema26.

No entanto, os desafios do agrupamento e da classificação que se manifestam no SISS-Geo vão além dos desafios clássicos destas tarefas.

Caracterização de um fenômeno.

A caracterização do que define um grupo de ocorrências (fenômeno) recai no problema da formulação de medidas de similaridade não convencionais (p.e., não necessariamente Euclidianas). Regras de agrupamento baseadas na experiência do especialista constituem uma alternativa razoável, mas esbarram na limitação da formalização do conhecimento e consequente potencial de introdução de vieses indesejáveis. Uma outra abordagem é tratar este problema como um processo de aprendizagem de máquina, objetivando o treinamento de modelos de similaridade: dado um novo registro e o conjunto de registrados existentes, determinar a qual grupo ele pertence—ou se caracteriza um novo grupo. O processo configura-se como aprendizado supervisionado, uma vez que é possível determinar confiavelmente, a priori ou a posteriori, quais registros pertencem a quais fenômenos, seja por exames laboratoriais ou convicção de especialistas.

Extração de características.

Uma vez constituídos os fenômenos é preciso avaliá-los quanto às potencialidades de ameaça à saúde silvestre e possível acometimento a humanos, pois, fenômenos por si só não configuram situações de alerta. Neste sentido, informações que caracterizam um grupo de ocorrências precisam ser extraídas e fornecidas ao modelo de predição de alerta. A dificuldade é, assim, derivar quais estatísticas melhor representam o fenômeno descrito pelo grupo a fim de maximizar o desempenho do modelo de predição; em outras palavras, levantar as informações que facilitem o processo de aprendizado. Especialistas preconizam o uso de certas estatísticas, como a espécie e quantidade de animais acometidos, número e frequência das ocorrências, entre outras; no entanto, o espaço de possíveis características vai muito além e poderia-se melhorar o desempenho preditivo. Dessa forma, uma questão em aberto é: como explorar esse vasto espaço automaticamente? Uma interessante linha de pesquisa e potencial solução para este desafio é a investigação de métodos de extração automática de características [10, 9].

[26]Afinal, é um sistema ambicioso que almeja agregar e hospedar os registros sobre saúde silvestre do vasto território nacional.

Modelo de predição de alerta.

Embora o seu uso no sistema aproxime-se de métodos suficientemente conhecidos e descritos na literatura, o modelo de predição de alerta é provavelmente o componente mais estratégico da inteligência do SISS-Geo. A viabilidade do sistema está fundamentalmente calçada no desempenho em termos de acurácia do modelo de predição, tanto na detecção de verdadeiro positivos (alertas) como de verdadeiro negativos (não alertas). A não detecção de uma situação de alerta (falso negativo) pode resultar em consequências graves à saúde silvestre, ambiental e, também, à humana. Por outro lado, os falsos positivos sobrecarregariam a escassa rede de laboratórios e especialistas responsáveis por confirmar ou negativar alertas (mais detalhes a seguir). Nesse sentido, métodos que combinam diversos modelos (ensemble methods) usualmente produzem soluções mais acuradas e robustas, sendo, portanto, candidatos promissores como algoritmos de treinamento dos modelos de predição [20]. Ainda, uma vez que a grande parcela de dados do sistema não possui classe associada, isto é, os fenômenos cujas predições de alerta ainda não foram confirmadas, o aprendizado semi-supervisionado constitui uma abordagem interessante em função da capacidade em também aproveitar instâncias não classificadas no processo de treinamento [3].

Confirmação de alerta.

Outro componente-chave do SISS-Geo—e do qual todos os demais dependem—é o processo de confirmação de alertas. O grande desafio e gargalo decorrem da necessidade da participação direta de humanos no procedimento de confirmação, seja em campo ou laboratório; portanto, é um processo caro e lento, mesmo considerando a extensa rede de colaborações qualificadas ligadas ao SISS-Geo. Quando há mais alertas emitidos pelo modelo de predição do que a capacidade de especialistas e rede de laboratórios em confirmá-los, os fenômenos precisam ser priorizados. Nesta situação, pode-se pensar em priorizar os fenômenos associados a alertas (1) pelo nível de alerta ponderado pela confiança da predição; ou (2) pela pertinência às regiões de grande interesse, seja este social, ambiental e/ou econômico. Entretanto, uma estratégia com enfoque em médio e longo prazo é a priorização da confirmação (ou negação) de alertas com maior potencial de aprimoramento da acurácia do modelo de predição. Esta linha de pesquisa é recente e denominada active learning [22]. Este mesmo método pode ser também empregado nos eventuais casos de falso negativos, evitando-se assim a possibilidade de degeneração do modelo de predição27: os fenômenos preditos como não alertas mas promissores sob o ponto de vista de aprendizado seriam passíveis de confirmação (da condição de não alerta) por especialista.

[27]Considere a situação extrema em que todas as predições são de *não alertas*, incluindo tanto verdadeiros quanto falso negativos. Dado que em princípio somente os casos de alertas são de interesse e passíveis de confirmação, neste cenário o modelo estaria fadado à degeneração.

3.3.2 Previsão de oportunidades ecológicas de ocorrência de doenças

Outra linha de fundamental importância no SISS-Geo é a previsão de cenários e ambientes que favoreçam oportunidades ecológicas para a ocorrência de doenças advindas da fauna silvestre ou, posto diferentemente, o levantamento de cenários propícios à ocorrência de certo evento, como por exemplo, um surto de uma determinada doença.

Resumidamente, os modelos de alerta treinados podem ser empregados para a avaliação de diversos cenários e caracterizar aqueles potencialmente suscetíveis. A construção dos modelos de previsão deverá relacionar diversas informações ambientais, sociais e de saúde humana e animal, mostrando-se uma área desafiadora aos atuais modelos de previsão. Métodos de relacionamento de variáveis ambientais e animais, tais como os aplicados na modelagem de nichos ecológicos ou mesmo os mais tradicionais métodos de aprendizagem de máquina serão amplamente aplicados nesse contexto, porém, novas abordagens devem ser elaboradas, permitindo a integração da variedade de informações citadas.

Além disso, devem-se considerar os desafios computacionais envolvidos na manipulação de informações de um grande número de registros, de diferentes espécies e condições ambientais, definindo-se como um problema de alto custo computacional. Entretanto, apesar da esperada manipulação de grandes massas de dados, espera-se também um reduzido número de informações sobre uma espécie ou doença específica, levando a um novo desafio: a aplicação de técnicas de predição em um ambiente altamente desbalanceado.

3.3.3 Extração de conhecimento

Uma propriedade importante dos métodos de modelagem simbólica, como árvores de decisão, algoritmos de extração de regras e a meta-heurística programação genética [14], é que o modelo é, ele próprio, a representação explícita do conhecimento extraído dos dados. Mais especificamente, é a revelação—passível de interpretação humana—das relações existentes entre os dados de entrada e saída.

É notável a potencialidade desta classe de modelos em assistir especialistas na análise e entendimento do fenômeno investigado, levando a interação homem-máquina curiosa: o modelo sugere hipóteses que melhor se ajustem aos dados enquanto o especialista as valida.

O desafio central da extração de conhecimento está na definição da estrutura/linguagem do modelo ou, em outras palavras, na incorporação do conhecimento do especialista. Nesse sentido, é também desafiador encontrar o balanço ideal entre viés, geralmente decorrente da simplicidade estrutural do modelo, e variância, questão normalmente associada aos modelos estruturalmente mais complexos.

Dependendo da parametrização dos algoritmos de aprendizagem e dimensionalidade da base de dados, um segundo desafio emerge: a demanda computacional associada, que é agravada pelo fato do processo de extração de conhecimento ser muitas vezes realizado interativamente pelo especialista. Tipicamente, as estratégias empregadas nestas situações incluem a (1) redução da dimensionalidade dos dados [8] e (2) computação paralela e distribuída, em arquiteturas convencionais ou aceleradores [1].

3.4 Ferramentas de Apoio à Rastreabilidade e à Composição de Análises sobre Saúde Silvestre

Ferramentas de análise e síntese de dados de biodiversidade a exemplo da modelagem de distribuição de espécies (MDE) [24], são amplamente utilizadas. Essas análises normalmente empregam diversas aplicações distintas, executadas de forma fracamente acoplada, caso típico para a utilização de sistemas de gerenciamento de workflows científicos [5]. Por exemplo, no caso da MDE, dados ambientais globais, como climatologia, uso da terra e topografia, são recuperados de provedores de dados ambientais, enquanto dados de ocorrências de espécies são obtidos de provedores como o GBIF. É comum que esses dados tenham que ser adaptados com sistemas de informações geográficas ou filtrados com ferramentas de controle de qualidade. Após esses passos de pré-processamento, algoritmos para MDE, a exemplo do Maxent [17], são aplicados para prever a distribuição potencial de espécies utilizando os dados ambientais e os de ocorrência de espécies adquiridos e manipulados nos passos de pré-processamento. Finalmente, uma etapa de pós-processamento é realizada, onde ferramentas estatísticas e de visualização de dados são utilizadas para analisar os dados resultantes da modelagem.

A utilização de sistemas de gerenciamento de workflows científicos permite que tais composições de diversas atividades, como ferramentas, sejam mais fáceis de especificar e executar por meio da automação de rotinas que frequentemente estão envolvidas nas mesmas:

- as atividades podem ter dependências de dados entre si, de modo que algumas atividades só podem iniciar a sua execução quando seus dados de entrada, produzidos por outras atividades que as precedem, estiverem disponíveis;

- eventualmente o fluxo de execução do workflow pode sofrer uma bifurcação para a execução de diversas atividades independentes entre si, tornando interessante a execução paralela das mesmas por questão de escalabilidade;

- o inicio da execução de uma atividade pode depender do fim da execução de diversas atividades disparadas após uma bifurcação, requerendo uma sincronização da execução do workflow, onde é necessário garantir que todas as atividades geradas pela bifurcação de fato terminaram a sua execução;

• caso o workflow utilize diversos recursos computacionais remotos, é necessário gerenciar a transferência de dados e monitorar a execução de atividades remotas.

Informações de proveniência [4], que reúnem detalhes sobre a concepção e a execução de processos computacionais, como por exemplo, workflows científicos, descrevendo os processos e dados envolvidos na geração dos resultados dos mesmos, podem ser utilizadas para facilitar esta tarefa. Elas permitem a descrição precisa de como o processo computacional foi projetado, chamada de proveniência prospectiva, e do que ocorreu durante sua execução, denominada proveniência retrospectiva. Algumas aplicações da proveniência incluem a reprodução de um processo computacional para fins de validação, compartilhamento e reutilização de conhecimento, verificação de qualidade de dados e atribuição de resultados científicos. Um dos conceitos comumente capturados na proveniência é o de causalidade, que é dado pelas relações de dependência existentes entre atividades computacionais e conjuntos de dados. Estas dependências podem derivar, por transitividade, dependências entre conjuntos de dados e entre processos. No contexto do SISS-Geo, as informações de proveniência permitirão que o processo de geração de alertas seja rastreável, ou seja, que seja possível recuperar os dados, parâmetros de configuração e atividades computacionais utilizados.

REFERÊNCIAS

[1] Augusto, D. A.; Barbosa, H. J. C. Accelerated parallel genetic programming tree evaluation with OpenCL. Journal of Parallel and Distributed Computing, Volume 73, Issue 1, Pages 86-100, 2012.

[2] Chame, M; Labarthe, N. Saúde Silvestre e Humana: Experiências e perspectivas. Fundação Oswaldo Cruz, Fiocruz, Rio de Janeiro, 2013.105p, 2013.

[3] Chapelle, O.; Schölkopf, B.; Zien, A. Semi-supervised learning. Cambridge, Mass.: MIT Press, 2006.

[4] Cuevas-Vicenttín, V.; Dey, S.; Köhler, S.; Riddle, S.; and Ludäscher, B. Scientific Workflows and Provenance: Introduction and Research Opportunities. Datenbank-Spektrum, 12(3):193–203, 2012.

[5] Deelman, E.; Gannon, D; Shields, M.; Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25(5):528–540, 2009.

[6] Estrada-Peña, A.; Ostfeld, R. S.; Peterson, A. T; Poulin, R.; Fuente, J. Effects of Environmental change on zoonotic disease risk: an ecological primer. Trends in Parasitology, 30(4):205-214, 2014.

[7] Fegraus, E. H.; Andelman, S; Jones, M. B.; Schildhauer, M. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. Bulletin of the Ecological Society of America, 86(3):158–168, 2005.

[8] Fodor, I. A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National, Technical Report, 2002.

[9] Guo, L.; Rivero, D.; Dorado, J.; Munteanu, C. R.; Pazos, A. Automatic feature extraction using genetic programming: An application to epileptic EEG classification. Expert Systems with Applications, 2011.

[10] Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. Feature Extraction, Foundations and Applications. Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[11] Hobern, D.; Apostolico, A.; Arnaud, E.; Bello, J. C.; Canhos, D.; Dubois, G.; Field, D.; García, E.; Hardisty, A.; Harrison, J.; Heidorn, B.; Krishtalka, L.; Mata, E.; Page, R.; Parr, C.; Price, J.; Willoughby, S. Global Biodiversity Information Outlook - Delivering Biodiversity Knowledge in the Information Age. Technical report, GBIF Secretariat, 2013.

[12] Jones, K.; Patel, N. G.; Levy, M. A.; Storeygard, A.; Balk, D.; Gitleman, J. L.; Daszak, P. Global trends in emerging infectious diseases. Nature, 451:990-993, 2008.

[13] Keesing, F.; Holt, R. D; Ostfeld, R. S. Effects of species diversity on disease risk. Ecology Letters, 9:485-495, 2006.

[14] Koza, J. R. Genetic programming: On the programming of computers by natural selection. MIT Press, Cambridge, Mass., 1992.

[15] Michener, W. K; Jones, M. B. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution, 27(2):85–93, 2012.

[16] Ostfeld, R; Glan, G. E.; Keesing, F. Spatial epidemiology: an emerging (or-reemerging) discipline. Trends in Ecology and Evolution, 20(6): 328-336, 2006.

[17] Phillips, S. J.; Anderson, R. P.; Schapire, R. E. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3-4):231–259, 2006.

[18] Poulin, R.; Forbes, M. Meta-analysis and research on host-parasite interactions: past and future. Evol. Ecol., 26:1169-1185, 2012.

[19] Poulin, R. Network analysis shinning light on parasite ecology and diversity. Trends in parasitology, 26:492-498, 2010.

[20] Rokach, L. Ensemble-based classifier. Artificial Intelligence Review, Volume 33, Issue 1-2, pp 1-39, 2010.

[21] Santos, S. D. R.; Huinca, S. C. M. Considerações sobre a utilização da PEC Padrão de Exatidão Cartográfica nos dias atuais. III Simpósio Brasileiro de Ciências Geodésicas e Tecnologias de Geoinformação. Recife, Pernambuco, 2009.

[22] Settles, B. Active Learning Literature Survey. Computer Sciences Technical Report, University of Wisconsin–Madison, 2009.

[23] Soberón, J.; Peterson, A. T. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 359(1444):689–98, 2004.

[24] Peterson, A. T.; Soberón, J.; Pearson, R. G.; Anderson, R. P.; Martínez-Meyer, E.; Nakamura, M.; Araújo, M. B. Ecological Niches and Geographic Distributions. Princeton University Press, 2011.

[25] Wieczorek, J.; Bloom, D.; Guralnick, R.; Blum, S.; Döring, M.; Giovanni, R.; Robertson, T.; Vieglais, D. Darwin Core: an evolving community-developed biodiversity data standard. PloS One, 2012.

[26] Xavier, S.D.C.; Roque, A.L.R.; Lima, V.S.; Monteiro, K.J.L., Otaviano, J.C.R. Lower Richness of small wild mammals species and Chagas disease risk. PLoS Neglected Tropical Diseases, 2012.

DESCOBERTA DE PADRÕES EM SISTEMAS URBANOS COMPLEXOS GESTÃO DA INFORMAÇÃO EM GRANDES VOLUMES DE DADOS MULTIMÍDIA DISTRIBUÍDOS (BIG DATA) EM TRANSPORTES

Ana L. C. Bazzan

Abstract. *Cities can be seen as networks that depict relationships among their inhabitants. For example, the relationship between the number of inhabitants and the number of schools, hospitals, fuel stations, etc. has been studied by re- searchers from the area of physics and complex systems. However, such studies tend not to address the issue about how such infra-structure is distributed in a city. One of the goals of this work is to investigate how techniques related to the field of networks can help public authorities to identify patterns arising in the area of transportation in particular and in the area of cities in general. One difficult step to use such techniques is the lack of data (in spite of the fact that the volume of such data tends to be huge and/or tend to increase). This happens because one part of such data is kept by public authorities, as well as because this data is distributed and heterogeneous. This way, a second goal of this work is data gathering and its analysis. Such data will be collected from heteroge- neous sources and is likely to involve natural language processing and/or image processing. Once this data is collected, the proposal is to perform an investiga- tion about properties of transportation networks, social networks among users of transit services, etc. in order to find out global patterns (comparison among cities), patterns at city level, as well as at local level (among different regions of a city).*

Resumo. Cidades podem ser vistas como networks que espelham relações en- tre seus habitantes. Por exemplo, foram estudadas relações entre o número de habitantes de várias cidades e o número de escolas, hospitais, postos de com- bustível, etc. Entretanto, tal estudo (e similares) raramente se ocupam com a questão de como a infra-estrutura esta distribuída em uma cidade. Um dos objetivos desta proposta é estudar como técnicas ligadas à área de networks podem auxiliar o poder público na identificação de padrões na área de trans- portes em particular e no âmbito de metrópoles em geral. Um dos passos para o uso destas técnicas é a dificuldade de se obter dados, apesar de seu volume ser, potencialmente, imenso. Isto se dá por que uma parte destes dados é de domínio do poder público e também por que os dados são heterogêneos e distribuídos. Desta forma, um outro objetivo é a coleta e tratamento de dados obtidos através de fontes heterogêneas como texto em linguagem natural e imagens. Uma vez de posse de tais dados, a proposta é realizar estudo sobre propriedades de redes de transporte, redes sociais entre usuários de transporte público, privado e de ride-sharing,

[1]PPGC / Instituto de Informática – UFRGS bazzan@inf.ufrgs.br

1PPGC / Instituto de Informática – UFRGS bazzan@inf.ufrgs.br

bem como de redes de infra-estrutura ligadas a transportes (taxis, lotações, postos de combustível), etc. a fim de descobrir padrões globais (com- paração entre cidades), em nível de uma cidade, e locais (em nível de regiões de cada cidade).

INTRODUÇÃO E MOTIVAÇÃO

Em 2006, a população urbana ultrapassou a rural em termos numéricos. De acordo com G. West[1], até 2050, um milhão de pessoas se agregam à cidades, *a cada semana*. Aglomerados urbanos atuam como atratores devido à diversidade de oportunidades em geral e à oferta de emprego em particular, o que está associado ao fato de que, cada vez mais, a economia mundial é calcada no setor de serviços e no setor industrial, ficando o setor primário em terceiro plano. Este fato está associado a muitos desafios e entender os padrões subjacentes é um passo importante para a solução de problemas ligados à urbanização.

Cidades podem ser vistas como *networks* ou seja como um grafo das interações físicas entre seus habitantes, resultando em agrupamentos de indivíduos, de infra-estrutura e de serviços, bem como em relações entre estes. Por exemplo, foram estudadas relações entre o número de habitantes de várias cidades e o número de escolas, hospitais, postos de combustível, etc. Em [Bettencourt et al. 2007, Kühnert et al. 2006] é mostrado que, no caso de infra-estrutura, esta relação está associada a uma lei de potências (power law) sub-linear o que significa uma economia de escala, ou seja, um incremento x na população está associado a um incremento $y < x$ na infra-estrutura existente. Por outro lado, alguns serviços (como restaurantes, cinemas) apresentam uma relação superlinear. Esta parece ser uma lei universal ou pelo menos se aplica a todas cidades estudadas.

Entretanto, tal estudo (e similares) raramente se ocupam com a questão de como a infra-estrutura esta distribuída em uma cidade. Como escolas públicas, hospitais, pontos de taxi e linhas de ônibus se distribuem nesta economia de escala? A sub-linearidade vale para todos os distritos de forma relativamente homogênea? Além dos aspectos de infra- estrutura, [Bettencourt et al. 2007, Kühnert et al. 2006] apontam outros, para os quais vale uma relação super-linear (salários, etc.). Assim sendo, como se dá a distribuição de salários, veículos particulares etc. nos diversos distritos da cidade?

Entender isto é talvez o maior desafio, uma vez que compreender como as leis de potências se aplicam em nível local possibilitaria o poder público agir no sentido de formular políticas públicas que viessem a corrigir desvios indesejáveis.

Um dos objetivos desta pesquisa é saber como técnicas ligadas à área de *networks* podem auxiliar o poder público na identificação de padrões na área de transportes em particular e no âmbito de metrópoles em geral. Neste texto foco na questão dos

trans- portes, por ter uma atuação maior nesta área. Um dos pontos importantes aqui é a questão de como um sistema eminentemente técnico pode se beneficiar dos seus usuários e seus dispositivos móveis quando estes atuam como sensores e estão conectados entre si (uma vez que tanto seres humanos quanto dispositivos fazem parte de redes sociais e, no futuro próximo, da Internet das coisas). Estes pontos são discutidos no artigo [Bazzan 2012] e no projeto "Inteligência Coletiva em Trânsito e Transportes: um Requisito para Cidades Inteligentes" (projeto vencedor do Prêmio Santander de Ciência e Inovação de 2012). O que é relevante aqui é que para se aplicar técnicas e ferramentas da área de *networks* é preciso obter dados.

Assim sendo, projetos atuais do grupo caminham no sentido de obter dados de outras fontes como a Internet. No artigo [Pereira et al. 2014] discutimos como uti- lizar dados de contexto para inferir o estado do trânsito. No projeto SAMU e em [Bazzan et al. 2013, Jorge L. Aching et al. 2014] nós propomos o uso de processamento de linguagem natural e de imagens a fim de obter dados de trânsito a partir de fontes het- erogêneas como Twitter, blogs e imagens de câmeras e de mapas que mostram o estado do trânsito de forma visual como o MapLink e Google traffic, além de informações visuais postadas pela CET (de SP). Outras possíveis fontes de dados são mapas de chamadas de celulares (para estimar quais são os distritos importantes como atrator e gerador de tráfego), informações sobre circulação de veículos de carga, dados sobre logística, bem como informações estáticas como a topologia das redes de transporte público e malha viária urbana. Em relação a isto, estamos trabalhando no tema de medidas de central- idade (como *closeness*, *betweenness*) em redes viárias [Galafassi e Bazzan 2013] e de transporte público (ver estudo sobre robustez da rede de metrô de S. Paulo [2]).

Uma vez de posse de tais dados, a proposta é realizar estudo sobre propriedades de redes de transporte, redes sociais entre usuários de transporte público, privado e de ride-sharing, bem como de redes de infra-estrutura ligadas a transportes (taxis, lotações, postos de combustível), etc. a fim de descobrir padrões globais (comparação entre cidades), em nível de uma cidade, e locais (em nível de regiões de cada cidade). Posteriormente, estender tal estudo para outros indicadores sócio-econômicos e realizar comparações e verificar leis de potências.

Espera-se poder colocar este projeto em prática em parceria com outros grupos de pesquisadores no Brasil e no exterior (atualmente existe uma colaboração com a FEUP, Portugal, através do Prof. R. Rossetti, com a Humboldt University, e com o SMART/MIT) e com o poder público e operadoras de celular interessadas em colaborar na detecção de padrões entre seus usuários, empresas de logística, além de outros provedores de informação.

[1]http://www.ted.com/talks/geoffrey_west_the_surprising_math_of_cities_and_corporations#t-722973

CONTEXTUALIZAÇÃO: PROBLEMA DOS TRANSPORTES EM REGIÕES METROPOLITANAS BRASILEIRAS

Um dos estopins das manifestações populares no Brasil em 2013 foi justamente a má qualidade dos transportes públicos no país. Embora o movimento tenha obtido um conge- lamento de tarifas, o problema persiste e só tende a piorar pois pouco foi feito em termos estruturais para melhorar o serviço.

No caso do transporte veicular em particular, no Brasil em geral e em S. Paulo em particular, congestionamentos são um problema do dia a dia[3]. De fato, a revista Times coloca São Paulo como tendo o pior trânsito do mundo[4].

Segundo o especialista Marcos Cintra em artigo publicado no jornal FSP, "o custo [...] é espantoso. Eles podem ser classificados em dois tipos: o tempo ocioso das pessoas no trânsito (conhecido em economia como "custo de oportunidade") e os gastos pecuniários impostos à sociedade.". No primeiro caso ele cita o aumento de R$ 15,4 bilhões (há 4 anos) para R$ 26,8 bilhões (2008) considerando somente o custo da hora não tra- balhada em São Paulo (apenas horários de pico). Soma-se a isto os gastos referentes ao consumo de gasolina e diesel, bem como o impacto dos poluentes na saúde da população e o aumento no custo do transporte de carga, cujo valor é estimado em R$ 6,5 bilhões por ano. Por fim o impacto negativo também se faz sentir na estrutura econômica da cidade de São Paulo e do país, na saúde das pessoas, no bolso do cidadão e na qualidade de vida da população.

"Soluções" como pedágios urbanos, rodizio de placas etc., praticadas atualmente no Brasil, são extremamente impopulares. O cidadão necessita ver o retorno do seu sac- rifício, seja ele monetário ou não. Desta forma, existe uma grande demanda por soluções que envolvam inteligência e informação como forma de oferecer uma contra-partida à população.

Os profissionais e técnicos que atuam na área de engenharia de transportes e tráfego há muito tempo já trabalham com ferramentas computacionais que os permitem estimar demandas e adequar a oferta de infraestrutura. Entretanto, estes profissionais não se valem de ferramentas (em parte por não existirem) que: (a) utilizam técnicas da área de networks; (b) atuam sobre grandes bases de dados; (c) se valem de geo-referenciamento.

Portanto, para o cientista da computação e o engenheiro de computação os desafios são inúmeros, em grande parte devido à complexidade dos problemas envolvidos.

[2]https://sites.google.com/site/trafficfactsfun/home/portugues/facts/transporte-publico/redes-de-metro
[3]344 km. em 23/05/2014: http://sao-paulo.estadao.com.br/noticias/geral,sao-paulo-atinge-maior-transito-de-sua-historia-com-344-km-parados, 1170756

MÉTODOS PROPOSTOS E PLANO DE AÇÃO

A metodologia de trabalho a ser empregada se concentra em três frentes.

Em primeiro lugar, pretende-se continuar a utilizar as ferramentas de simulação microscópica e baseadas em agentes para investigar fenômenos macroscópicos como ocorrência de congestionamentos em função de modificações na rede viária e/ou mod- ificações na políticas de uso e controle sobre a rede viária. Para tanto serão utilizadas ferramentas como o SUMO[5] [Behrisch et al. 2011, Krajzewicz et al. 2012] ou outras que foram desenvolvidas no grupo [Silva et al. 2006, Bazzan et al. 2011].

Com tal, objetiva-se dois aspectos. Primeiro, a implementação, validação e comparação de modelos que estão sendo propostos no âmbito do grupo de pesquisa e de colaborações externas, voltados às áreas de controle semafórico, alocação de viagens e outras.

O segundo aspecto, que é mais relevante para o contexto deste trabalho, é a in- vestigação, em ambiente simulado, do que ocorre quando se tem uma certa topologia da rede de transporte, bem como uma certa distribuição de objetos e serviços na rede que representa uma cidade.

Até este momento, paradigmas de simulação baseada em agentes têm focado em questões específicas e locais como responder a perguntas sobre como age um agente motorista em cenários que basicamente abrangem escolha de rotas (uma forma de *traffic assignment*). Um número menor de trabalhos lidam com modelagens de outros com- ponentes do sistema viário através de agentes. Outras possibilidades seriam agentes semáforo, agentes pedestres, e agentes veículos. Para uma visão geral sobre este tema, ver [Bazzan e Klügl 2013]. Entretanto, poucos trabalhos lidam com ambientes onde to- dos estes agentes interagem ativamente e como estes agentes interagem com o ambi- ente quando tem informação sobre sua estrutura topológica, sobre centralidade de partes da malha viária, e sobre o estado de determinados atributos (por exemplo existência ou não de congestionamentos). Este é um dos focos deste projeto. Além disto, no caso de agentes veículos, com a tendência de se ter a chamada *Mobility Internet* (um conceito lançado no livro "*Reinventing the Automobile*", [Mitchell et al. 2010]), os veículos trocarão informação entre si ou, no mínimo, disponibilizarão informações diversas, assim como smartphones o fazem atualmente. Tratar estes dados de forma eficiente vai exigir novos paradigmas de recuperação de informação, simulação e otimização.

[4]"The World's Worst Traffic Jams", Times Magazine. Acessado em 21 Dez. 2008 em http://www. time.com/time/world/article/0,8599,1733872,00.html
[5]http://sumo.sourceforge.net

Ainda que atualmente a *Mobility Internet* não seja uma realidade concreta, os sistemas avançados de informação ao motorista (e.g., [Bonsall 1992, Bonsall et al. 1997, Dia e Purchase 1999, Dia e Panwai 2014]), onde os usuários do sistema viário poderiam ter acesso a uma série de informações que já se encontram parcialmente disponíveis (at- ualmente apenas na Internet). Isso pode ser alcançado aumentando a interconectividade dos componentes do sistema viário através do uso de técnicas da área de sistemas de in- formação e comunicação (como por exemplo Advanced Traveller Information System), bem como com comunicação inter-veicular, a chamada comunicação veículo a veículo ou interveicular, veículo a infraestrutura, etc.

É comprovado [Yang et al. 2010] que a troca de informação entre veículos pode ajudar a melhorar a distribuição destes na rede de transporte, diminuindo congestionamen- tos e tempo de percurso. O que está em aberto no momento é exatamente como isto pode ser feito i.e. usando que tipos de protocolos, etc. Um primeiro passo neste sentido aparece em [Rybicki 2011, Scheuermann et al. 2009, Lochert et al. 2008, Rybicki et al. 2007].

Em função do fato de que validar tais modelos não é atualmente trivial dada a falta de dados reais, isto motivou o direcionamento da pesquisa para a obtenção de dados de fontes diversas. Portanto, como segunda frente, este projeto visa o uso de técnicas de recuperação de informações heterogêneas oriundas de fontes como Web e comunicação inter-veicular, a fim de alimentar sistemas de tomada de decisão e informação ao usuário. Pretende-se continuar os trabalhos nas áreas de coleta de informação sobre trânsito, a par- tir de fontes heterogêneas. Neste rol encontram-se fontes usuais como dados de sensores convencionais (laço induzido), câmeras, bem como novas fontes de dados como dados so- bre trânsito e tráfego na Internet, por meio de blogs oficiais, páginas de eventos esportivos e culturais, etc. Este trabalho apresenta desafios técnicos importantes uma vez que se trata de informação não estruturada, normalmente não geo-referenciada, e em linguagem nat- ural para a qual em geral não dispomos de um corpus anotado. Uma outra fonte de dados advem de informação trocada entre veículos através da chamada comunicação interveic- ular. Como este tipo de dado ainda não está disponível, utiliza-se ambientes simulados, como o SUMO anteriormente mencionado.

Por fim, uma vez que a coleta de informações sobre trânsito e tráfego gerará pos- sivelmente uma grande quantidade de dados, é possível se aplicar conceitos e métricas da área de *networks* a fim de detectar padrões interessantes. Ressalta-se que alguns destes da- dos necessitam ser, necessariamente, geo-referenciados. Por isso, estamos desenvolvendo métodos para se geo-referenciar os dados coletados.

Uma primeira tentativa está sendo realizada através do estudo de redes de transporte público. Especificamente, estão sendo realizados cálculos sobre a robustez do siste-

ma metroviário da cidade de S. Paulo (baseado em estudos simi- lares realizados sobre redes metroviárias de outras cidades, [Derrible e Kennedy 2009, Derrible e Kennedy 2010]) utilizando-se medidas de centralidade. Este estudo tam- bém está sendo realizado para malhas viárias em geral, estendendo trabalhos do grupo [Galafassi e Bazzan 2013].

Cabe ressaltar que esta terceira linha tem grande potencial para crescer uma vez que cada vez mais veremos a produção de um volume crescente de dados gerados não apenas pelo fenômeno da *Mobility Internet*, como também pela disseminação de dados produzidos por dispositivos móveis, dentre os quais uma parte considerável refere-se a temáticas ligadas à cidade em geral e a transportes em particular.

Desta forma, a metodologia de trabalho a ser empregada consiste em utilizar téc- nicas de sistemas complexos, *networks*, inteligência artificial e ferramentas de simulação microscópica e baseadas em agentes. Assim, espera-se tirar conclusões que possam auxiliar não apenas o poder público mas também, e principalmente, os diversos atores do sistema viário (motoristas, passageiros, usuários de outros sistemas como car e ride shar- ing, empresas de logística, etc.).

No que se refere à recuperação de informação de fontes heterogêneas, pretende-se dar continuidade aos trabalhos em andamento no grupo como por exemplo recuperação de informação textual de redes sociais, e coleta de informações provenientes de comunicação inter-veicular (IVC). No caso de IVC, temos parceria com o grupo de pesquisa do Prof. Scheuermann da Univ. Humboldt, Alemanha. O objetivo desta interação é somar esforços na área de simulação de tráfego com foco em IVC. O grupo alemão é especialista na área de redes de dados para este tipo de aplicação, enquanto que o grupo na UFRGS é especialista em simulação baseada em agentes.

No que tange a questão de uso de informações textuais, temos colaborações infor- mais com pesquisadores como Francisco Pereira (MIT) e Rosaldo Rossetti (FEUP).

Em linhas gerais, para levar adiante as três frentes mencionadas, é proposto o seguinte plano de trabalho.

- Atualização bibliográfica sobre os temas abordados;

- Estudo e desenvolvimento de métodos para extração de informação de imagens (e.g., sobre estado do trânsito como fornecida atualmente por sites como MapLink e Google traffic);

- Continuação do desenvolvimento de métodos para extrair e processar informações textuais (e.g., provenientes de blogs e Twitter sobre a temática trânsito);

- Continuação do desenvolvimento de métodos para processar IVC;

- Definição sobre o cenário para uso de IVC;
- Estudo sobre uso de outras fontes de dados;
- Desenvolvimento de mecanismo para extração e disponibilização de informações sobre o trânsito a partir de fontes heterogêneas;
- Desenvolvimento de mecanismo de apoio à tomada de decisão dos usuários com base nas informações extraídas no item anterior;
- Definição das plataformas de simulação a serem utilizadas;
- Desenvolvimento de camada de modelagem baseada em agentes para o simulador escolhido;
- Definição das redes viárias de teste;
- Desenvolvimento de algoritmos de escolha de rota;
- Desenvolvimento de algoritmos de controle semafórico;
- Extensão de algoritmos de escolha de rota para utilizar informações de fontes heterogêneas sobre trânsito;
- Extensão de algoritmos de escolha de rota para permitir a cooperação entre os agentes por meio de troca de informações;
- Etapa de extração e processamento de texto em linguagem natural;
- Etapa de geo-referenciamento de informações obtidas de fontes não geo-referenciadas;
- Etapa de aplicação de métodos ligados à área de *networks* (aqui pretende-se ini- ciar analisando dados que sejam o mais diretamente possível representados sob a forma de grafos);
- Aplicação de métodos como medidas de centralidade para obter padrões em áreas urbanas, relacionados com infra-estrutura e serviços ligados à área de transportes;
- Estudos comparativos sobre estes padrões, como por exemplo comparação entre cidades, entre áreas de uma cidade e eventualmente localmente (sub distritos);
- Análise e comparação dos métodos acima gerando recomendações para o futuro desenvolvimento de cidades inteligentes;
- Estudo do impacto das técnicas propostas;
- Documentação e redação de relatórios técnicos, artigos e materiais didáticos.

CONTRIBUIÇÕES PRETENDIDAS

Considerando os aspectos desta pesquisa, os impactos esperados são:

• Impacto socioeconômico:

- Contribuição para a definição de métodos mais eficientes de utilização da malha viária, reduzindo os custos bilionários de sua ampliação

- Redução de estresse dos usuários do sistema de mobilidade urbana através da redução no número de horas paradas no trânsito

- Redução dos custos de oportunidade

- Melhoria na percepção dos usuários do potencial logístico do país

- Formação de recursos humanos

• Impacto técnico-científico:

- Novos resultados quanto à descoberta de padrões em sistemas urbanos

- Novos resultados quanto à modelagem de tráfego baseada em agentes

- Novos resultados quanto à modelagem de motoristas sob os aspectos com- portamental e informacional

- Novos resultados quanto ao impacto da extração e uso de informações no processo decisório dos agentes

- Novos métodos de cálculo de rotas de forma distribuída e utilizando co- municação inter-veicular

- Disponibilização de algoritmos eficientes para: planejamento de rotas, controle semafórico e extração de informações

- Acúmulo de experiência quanto à aplicabilidade de técnicas ligadas a agentes inteligentes em cenários de tráfego de larga escala

- Descrição analítica dos efeitos da comunicação inter-veicular e de técnicas de aprendizagem de máquina

- Descrição analítica dos resultados dos estudos utilizando técnicas da área de *networks*

- Integração dos algoritmos de roteamento em uma ferramenta de simulação microscópica, permitindo analisar aspectos de demanda e de controle

- Publicações em conferências e periódicos, nacionais e internacionais, de alta relevância e impacto

AGRADECIMENTOS

Agradeço à SBC pela organização do Seminário Grandes Desafios 2014, bem como ao CNPq (pelo auxílio concedido no Edital Universal 2012, auxílio no âmbito da bolsa

de produtividade e pesquisa, e auxílio no âmbito da cooperação internacional com a Alemanha/BMBF/DLR), à Fundação Alexander von Humboldt (fellowship), FAPERGS (PRONEX), Fundo Setorial de Transportes (edital de 2009 conjunto com o CNPq), Santander (prêmio de Ciência e Inovação 2012), CAPES (bolsas no âmbito do PNPD), e ao ITL/SENAT (bolsas de pós graduação no âmbito do projeto SAMU).

REFERENCES

Bazzan, A. L. C. (2012). Lessons learned from one decade of developing agent-based tools for traffic modeling, simulation, and control: how to make cities smarter. In *Braz. Symp. on Information Systems (SBSI)*, pp. 67–72. SBC.

Bazzan, A. L. C., Amarante, M. d. B. do., Azzi, G. G., Benavides, A. J., Buriol, L. S., Moura, L., Ritt, M. P. e Sommer, T. (2011). Extending traffic simulation based on cellular automata: from particles to autonomous agents. In Burczynski, T., Kolodziej, J., Byrski, A. e Carvalho, M., editores, *Proc. of the Agent-Based Simulation (ABS / ECMS 2011)*, pp. 91–97, Krakow. ECMS.

Bazzan, A. L. C., Araújo, P. G., Galafassi, C., Tavares, A. R., Vecchia, A. D., de Vit, A.

R. D. e Vivian, G. R. (2013). Smart drivers: Simulating the benefits of giving twitter information about traffic status. In *Anais do Congresso da SBC 2013*, pp. 249–259. SBC.

Bazzan, A. L. C. e Klügl, F. (2013). A review on agent-based technology for traffic and transportation. *The Knowledge Engineering Review*, 29(3):375–403.

Behrisch, M., Bieker, L., Erdmann, J. e Krajzewicz, D. (2011). SUMO - simulation of urban mobility: An overview. In *SIMUL 2011, The Third International Conference on Advances in System Simulation*, pp. 63–68, Barcelona, Spain.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. e West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306.

Bonsall, P., Firmin, P., Anderson, M., Palmer, I. e Balmforth, P. (1997). Validating the results of a route choice simulator. *Transportation Research C*, 5:371–387. Elsevier, Oxford.

Bonsall, P. W. (1992). The influence of route guidance advice on route choice in urban networks. *Transportation*, 19(1).

Derrible, J. S. e Kennedy, C. (2009). A network analysis of subway systems in the world using updated graph theory. *Transportation Research Record*, 2112:17–25.

Derrible, S. e Kennedy, C. (2010). The complexity and robustness of metro networks.

Physica A: Statistical Mechanics and its Applications, 389:3678–3691.

Dia, H. e Panwai, S. (2014). *Intelligent Transport Systems: Neural Agent (Neugent) Models of Driver Behaviour*. LAP Lambert Academic Publishing.

Dia, H. e Purchase, H. (1999). Modelling the impacts of advanced traveller information systems using intelligent agents. *Road and Transport Research*, 8:68–73.

Galafassi, C. e Bazzan, A. L. C. (2013). Analysis of traffic behavior in regular grid and real world networks. In *The Fifth International Workshop on Emergent Intelligence on Networked Agents (WEIN)*.

Jorge L. Aching, S., de Oliveira, T. B. F. e Bazzan, A. L. C. (2014). Traffic information extraction from a blogging platform using a bootstrapped named entity recognition approach. In *IEEE Symposium Series on Computational Intelligence (SSCI 2014)*, Orlando. IEEE. to appear.

Krajzewicz, D., Erdmann, J., Behrisch, M. e Bieker, L. (2012). Recent development and applications of SUMO - Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements*, 5(3&4):128–138.

Kühnert, C., Helbing, D. e West, G. B. (2006). Scaling laws in urban supply networks.

Physica A, 363:96–103.

Lochert, C., Rybicki, J., Scheuermann, B. e Mauve, M. (2008). Scalable data dissemination for inter-vehicle-communication: Aggregation versus peer-to-peer. *Oldenbourg IT – Information Technology*, 50(4):237–242.

Mitchell, W. J., Borroni-Bird, C. E. e Burns, L. D. (2010). *Reinventing the Automobile*.

MIT Press, Cambridge, MA.

Pereira, F. C., Bazzan, A. L. C. e Ben-Akiva, M. (2014). The role of context in transport prediction. *IEEE Intelligent Systems Magazine*, 29(1):76–80. ITS Department.

Rybicki, J. (2011). *Cooperative Traffic Information Systems Based on Peer-to-Peer Networks*. Tese de Doutorado, Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf.

Rybicki, J., Scheuermann, B., Kiess, W., Lochert, C., Fallahi, P. e Mauve, M. (2007). Challenge: Peers on wheels – a road to new traffic information systems. In *Mobi- Com '07: Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, pp. 215–221.

Scheuermann, B., Lochert, C., Rybicki, J. e Mauve, M. (2009). A fundamental scalability criterion for data aggregation in VANETs. In *MobiCom '09: Proceedings of the*

15th Annual ACM International Conference on Mobile Computing and Networking, pp. 285–296.

Silva, B. C. da., Junges, R., Oliveira, D. e Bazzan, A. L. C. (2006). ITSUMO: an intelligent transportation system for urban mobility. In Nakashima, H., Wellman, M. P., Weiss, G. e Stone, P., editores, *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pp. 1471–1472. ACM Press.

Yang, Y., Li, X., Shu, W. e Wu, M.-Y. (2010). Quality evaluation of vehicle navigation with CPS. In *GLOBECOM '10: Proceedings of the IEEE Global Communications Conference*.

GESTÃO DA INFORMAÇÃO EM GRANDES VOLUMES DE DADOS MULTIMÍDIA DISTRIBUÍDOS (BIG DATA) NO SISTEMA BANCÁRIO/FINANCEIRO

Claudio S. Pinhanez

Abstract. This paper explores the contributions and challenges related to the use of social data to improve and evolve the banking and financial system. Starting with some examples of the use of social media and network data, money transactions, and other sources of social data, we argue their huge transformational power in the development of new financial systems with improved efficiency, productivity, and social inclusion. However, to realize this potential requires the development of new technologies not only to handle large volumes of data but also to deal with issues such as privacy and security of social, computational modeling of financial life of individuals and small enterprises, digital money and payments, bank branch transformation, and financial education. The paper concludes by listing scientific challenges for computer science in the context of the development of those new technologies for the financial and banking sectors.

Resumo. Este artigo explora as contribuições e desafios do uso de dados das relações sociais de indivíduos na melhoria e evolução do sistema bancário e financeiro. A partir da exploração de exemplos de usos de dados de mídias e redes sociais, transações monetária, e outras fontes de dados sociais, argumenta-se o enorme poder transformacional desses dados na construção de novos sistemas financeiros com maior eficiência, produtividade e inclusão social. Contudo, a realização desse potencial exige o desenvolvimento de novas tecnologias não só para lidar com os grandes volumes de dados mas também para lidar adequadamente com questões como privacidade e segurança de dados sociais, modelamento computacional da vida financeira de indivíduos e pequenas empresas, dinheiro e pagamento digitais, transformação de agências bancárias e educação financeira. O artigo conclui listando alguns desafios científicos para a área de ciência da computação dentro do contexto do desenvolvimento dessas novas tecnologias para o setor financeiro e bancário. Ao longo dos últimos dois anos o grupo de pesquisa em Analítica de Dados Sociais da IBM Research – Brasil tem se dedicado a projetos de pesquisa que envolvem diretamente o uso de Big Data no contexto do sistema bancário e financeiro. Entre outros, este grupo de pesquisa tem explorado usos de vários tipos de Big Data, incluindo mídias sociais, redes sociais,

IBM Research – Brasil
Rua Tutóia, 1157, São Paulo, SP, CEP 04007-900

csantosp@br.ibm.com

transações financeiras e avaliações de crédito no contexto de novos sistemas na área bancária e financeira.

Exemplos de projetos de pesquisa na área incluem sistemas de detecção de eventos de vida e oportunidades de negócio a partir do processamento de textos em mídias sociais;.o uso das informações de interconexão pessoal, coletadas em sites de redes sociais, para a determinação de avaliação de crédito, especialmente da população desbancarizada e da economia informal; e o uso de redes sociais no estabelecimento de sistemas de pagamento e de dinheiro digital seguros e de baixo custo transacional.

Além da experiência proporcionada por esses projetos de pesquisa, membros do grupo tem dialogado constantemente com profissionais de TI do setor bancário e de seus fornecedores, explorando exaustivamente as oportunidades e desafios para o uso de Big Data no setor bancário e financeiro. Em 2013, o grupo participou ativamente de um workshop de 2 meses com um banco emergente, interagindo diariamente com os setores de TI, de experiência do consumidor e de marketing do banco, definindo um roadmap para o estabelecimento de um banco digital inovador baseado na utilização maciça de informações estruturadas e não estruturadas. Ao fim deste trabalho, não só foi obtida uma visão sistemática das diferentes tecnologias disponíveis no mercado nessa área, como também foi possível determinar quais as áreas prioritárias de pesquisa e desenvolvimento para a criação de uma nova experiência bancária.

Também em 2013 membros do grupo de Analítica de Dados Sociais da IBM Research – Brasil participaram em um workshop de inovação com um dos maiores bancos do Brasil, no qual houve a participação da grande maioria dos diretores e executivos do banco, onde foram exploradas e discutidas as principais oportunidades de inovação no setor bancário, incluindo as áreas de pagamentos, dinheiro digital, agências, marketing, e inclusão social. Nestas discussões várias aplicações, existentes e desejáveis, de Big Data foram exploradas em termos de impacto, relevância, e capacidade de alavancagem. Este workshop de inovação resultou em um processo de elaboração de projetos pioneiros de transformação bancária no Brasil no qual o grupo está engajado.

Nos últimos 12 meses o grupo também esteve envolvido em várias projetos com as maiores instituições financeiras do país diretamente relacionadas ao uso e potencial impacto de infraestruturas de Big Data, especialmente dados sociais, em negócios bancários. Em particular, foram desenvolvidas tecnologias, baseada em sistemas de aprendizado de máquina, para a detecção de oportunidades de negócios bancários a partir de postagens públicas de clientes em sites de mídia social. Como parte desses projetos, vários milhões de postagens foram processadas e classificadas, com alta precisão e acurácia, em termos de sua potencial utilidade como oportunidades de negócios.

O desenvolvimento deste tipo de tecnologia continua sendo feito na IBM Research – Brasil, hoje um dos laboratórios no país com maior experiência de processamento e análise de grandes volumes de dados de redes sociais. Em projetos paralelos, o grupo foi responsável pelo desenvolvimento de tecnologia baseada em métodos Deep Learning para processamento em tempo real de postagens sobre futebol. Esta tecnologia foi utilizada durante a Copa do Mundo para o processamento de todos as postagens em português durante os 64 jogos, totalizando mais de 53 milhões de postagens analisadas em tempo real.

A partir dessas experiências diversas, o grupo tem identificado vários desafios importantes, bem como oportunidades de novas tecnologias, para o uso efetivo de Big Data, e especialmente dados de mídias e redes sociais, no setor financeiro e bancário. Além disso, o grupo tem explorado e investigado o uso de dados públicos, estruturados e não estruturados, no suporte à inclusão de desbancarizados e de agentes da economia informal no ecosistema bancário. Através do uso dessas informações parece ser possível uma melhoria da avaliação de risco e confiabilidade de novos clientes, permitindo que indivíduos e pequenas empresas possam abrir contas e receber crédito mesmo sem histórico bancário e financeiro.

Outra importante área de pesquisa do grupo de Análitica de Dados Sociais do laboratório da IBM Research no Brasil é em dinheiro e pagamentos digitais, e em especial no emprego de redes sociais (baseadas em sites tradicionais ou não) para o estabelecimento de formas simples, seguras, e baratas de sistemas de pagamentos. De maneira similar, o grupo está estudando sistemas de microcrédito social, hoje baseados em relações face-a-face, e em como expandir e facilitar o acesso ao microcrédito usando informações contidas em grandes volumes de informações.

Vários desses projetos, seus resultados e as tecnologias desenvolvidas foram e estão sendo publicados em artigos científicos em jornais e conferências, permitindo uma troca de informações, idéias, e experiência efetiva com a comunidade acadêmica. Como parte da proposta de participação do grupo de Análitica de Dados Sociais do laboratório da IBM Research no Brasil no 3º Seminário dos Grandes Desafios de Computação, através do seu gerente e pesquisador sênior Claudio S. Pinhanez (vide resumé no Anexo 1), é proposta a escrita e apresentação de um artigo que reporte as necessidades, oportunidades e dificuldades do uso de dados sociais no setor bancário e financeiro, a partir da extensa interação com profissionais e executivos que o grupo teve ao longo dos último anos.

O artigo a ser submetido na segunda fase de seleção, provisoriamente entitulado *"Contribuições e Desafios de Tecnologias de Dados Sociais na Expansão e Melhoria do Sistema Bancário e Financeiro no Brasil: Eficiência, Produtividade, e Inclusão Social"*, abordará também as limitações das tecnologias de Big Data em uso e os desafios

principais para a Ciência da Computação nesse contexto. O anexo 2 contém uma proposta de página de rosto para essa contribuição, incluindo uma primeira versão do resumo do artigo.

Poucos grupos de pesquisa no Brasil tem o grau de exposição e interação que a IBM Research – Brasil tem com os profissionais e técnicos do setor bancário e financeiro. Este nível de envolvimento com o setor bancários é em parte obtida pela posição da IBM de principal fornecedor de infraestrutura e software no mercado brasileiro. Além disso, o grupo de Análitica de Dados Sociais tem desenvolvido vários projetos pioneiros de uso de Big Data em diversas áreas do setor financeiro, relatados brevemente aqui, e que podem fornecer uma perspectiva mais realista a respeito da importância relativa de diferentes sub-áreas de Big Data no contexto do setor financeiro e bancário.

Em resumo, a participação de Claudio S. Pinhanez no 3º Seminário dos Grandes Desafios da Computação, na área de Big Data no Sistema Bancário e Financeiro, permitirá contribuições a partir de uma perspectiva quase única no Brasil, de um grupo de pesquisadores de alto nível em computação que estão efetivamente em permanente e constante contato com profissionais da indústria, e com experiência prévia em projetos de Big Data no setor.

DESAFIOS E OPORTUNIDADES EM NEUROCIÊNCIA COMPUTACIONAL NA EDUCAÇÃO BRASILEIRA

Raimundo José Macário Costa[1], Luís Alfredo Vidal de Carvalho[2], Emilio Sánchez Miguel[3], Renata Mousinho[2], Renato Cerceau[2,5], Lizete Pontes Macário Costa[4], Sérgio Manuel Serra da Cruz[1,6]

Resumo. Compreender o funcionamento do cérebro se constitui em um dos grandes desafios do momento. Nas áreas das neurociências e da educação, novos estudos buscam correlacionar as dificuldades de aprendizagem enfrentadas por crianças e jovens com problemas comportamentais e sociais. Este trabalho tem por objetivo apresentar os desafios e oportunidades da pesquisa em neurociência computacional com vistas a detectar pessoas com Transtornos de Aprendizagem. Apresentamos uma linha de investigações baseadas em redes neurais que considera jovens na faixa etária de 9 a 18 anos com ou sem o Transtorno de Aprendizagem. A adoção das redes neurais demonstra consistência ao lidar com os problemas de reconhecimento de padrões e se mostram eficientes na detecção precoce em portadores destes transtornos.

1. DESCRIÇÃO DO PROBLEMA

A compreensão do funcionamento do cérebro ainda permanece como um dos grandes desafios no meio científico no século XXI (ABBOT, 2013). As pesquisas sobre o tema vêm crescendo de modo exponencial desde os anos 1960. A Neurociência é uma área mais recente e também passou a crescer de modo significativo a partir de dos anos 1980 (Figura 1). A Neurociência tem como objetivo estudar e analisar o sistema nervoso central (SNC) dos seres humanos e animais, suas funções, formato particular, fisiologia, lesões ou patologias. Esta área conseguiu importantes avanços que promoveram efeitos positivos sobre a qualidade de vida dos pacientes que sofrem, por exemplo, da Esclerose Múltipla, de Doença de Alzheimer, de Parkinson e outras doenças relacionadas ao SNC (LENT, 2001). No entanto, apesar dos grandes investimentos na área, muito ainda está por se realizar, especialmente no que diz respeito à compreensão dos mecanismos de ligação entre as estruturas cerebrais e a

[1]Universidade Federal Rural Rio de Janeiro (UFRRJ)
[2]Universidade Federal Rio de Janeiro (UFRJ)
[3]Universidad de Salamanca (USAL)
[4]Universidade do Estado do Rio de Janeiro (UERJ)
[5]Agência Nacional de Saúde Suplementar (ANS)
[6]Programa de Educação Tutorial (PET-SI/UFRRJ)

{macario, serra}@ufrrj.br

funcionalidades em nível microscópico aos processos cognitivos e comportamentais (MARKRAM, 2013).

No início dos anos 90, na área científica foi proferida a "década do cérebro". Esta denominação teve origem nos EUA e buscava incentivar a identificação dos processos neuropsicobiológicos normais e os distúrbios relacionados. Neste contexto, aliado aos expressivos avanços da Ciência da Computação e a disseminação da rede Internet a área de Neurociência Computacional floresceu (SCHWARTZ, 1990). Desde então vem desenvolvendo esforços e buscando novas estratégias para o desenvolvimento de modelos matemáticos e computacionais realísticos a fim de simular o cérebro.

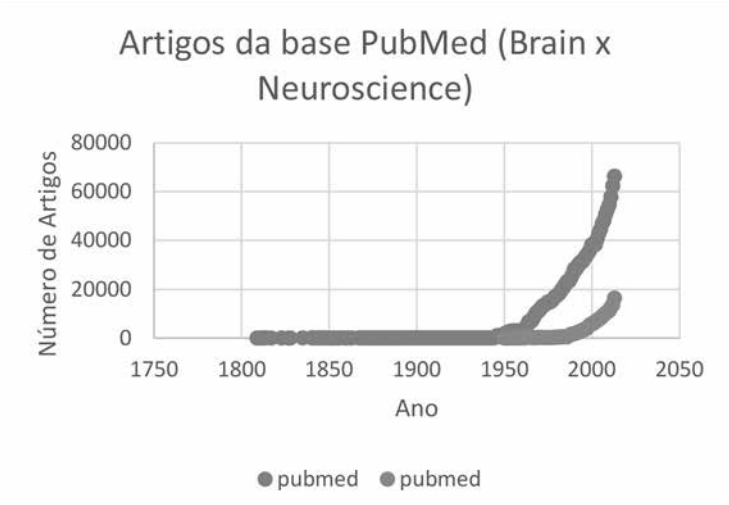


Figura 1. Comparativo entre quantitativo dos artigos publicados na base PubMed incluindo os termos "brain" e "neuroscience" no título (Setembro, 2014).

Mais recentemente em abril de 2013, foram reapresentados nos EUA e na Europa grandes projetos de investigação denominados *BRAIN Initiative* (NIH, 2014) e *Human Brain Project* (HBP, 2014) respectivamente. As iniciativas reapresentam demandas que se propõem a revolucionar a compreensão do funcionamento dos mistérios do cérebro humano buscando acelerar o desenvolvimento de novas tecnologias que permitirão aos pesquisadores e cientistas obterem imagens dinâmicas do cérebro em ação, exibindo como células cerebrais e complexos circuitos neurais interagem na velocidade do pensamento, ampliando a base de conhecimento sobre como pensamos, aprendemos e lembramos.

A estrutura proposta para a realização da iniciativa BRAIN inclui empresas privadas, centros de pesquisa e agências governamentais além de um grande leque de espe-

cialistas que vão desde Médicos, Neurocientistas, Nanocientistas, Bioinformatas aos Engenheiros e Cientistas da Computação, em especial aqueles que atuam nas áreas de Inteligência Artificial, Bancos de Dados, Computação Gráfica, HPC, *Big Data*, *E-Science*, *Web*, Jogos, Robótica, Sensores, Redes Sociais entre outros (ZHONG, 2012; NIH, 2014; HBP, 2014).

No Brasil registram-se atuações significativas de neurocientistas em prol do desenvolvimento do conhecimento relacionado ao cérebro e da "Indústria Brasileira do Cérebro". Dentre os diversos centros de pesquisas podemos citar os trabalhos desenvolvidos no Instituto Internacional de Neurociências de Natal Edmond e Lily Safra (IINN-ELS) e no Instituto de Ciências Biomédicas da UFRJ (ICB-UFRJ). Apesar destes esforços e da sinergia entre a Neurociência Computacional e a Educação, o Brasil ainda carece de maiores estudos que correlacionam os Transtornos de Aprendizagem enfrentados por crianças e jovens com diversas técnicas computacionais. Tais questões tem profunda relevância social e podem repercutir na evasão escolar, analfabetismo funcional e sucessivas reprovações.

O objetivo deste artigo é expor direções de pesquisa e considerações sobre os desafios referentes ao apoio computacional oferecido pela Neurociência Computacional na Educação Brasileira. O trabalho também apresenta uma abordagem relacionada com a dislexia, um dos Transtornos de Aprendizagem que tem despertado interesse nos pesquisadores, profissionais de saúde e escolas. A dislexia pode ser caracterizada por uma falha no processo de aquisição e/ou desenvolvimento das habilidades escolares.

2. ESCOPO DA SOLUÇÃO

A neurociência computacional pode ser empregada para a construção de um sistema computacional inteligente capaz de tratar e analisar grandes volumes de dados semiestruturados, elaborar jogos educacionais ou mesmo desenvolver aplicações móveis direcionadas ao apoio diagnóstico e rastreio de Transtorno de Aprendizagem.

A Neurociência Computacional é essencialmente interdisciplinar e assim está apoiada em três pilares: *Neurociência* (das áreas de Medicina e Ciências Biológicas); *Ciência Cognitiva* (da área de Psicologia) e *Ciência da Computação* (áreas de Inteligência Artificial, Bancos de Dados, *E-Science*, Proveniência, *Big Data*, *High Performance Computing (HPC)*, Nuvens, entre outros) (Figura 2). No entanto, uma das grandes dificuldades desta área é modelar (matemática e computacionalmente) um Transtorno de Aprendizagem, identificar quais são dados e variáveis mais relevantes, transcrevê-las para as soluções tecnológicas e avaliar se os resultados computacionais são significativos e válidos de acordo com os aspectos médico, ético e educacional.

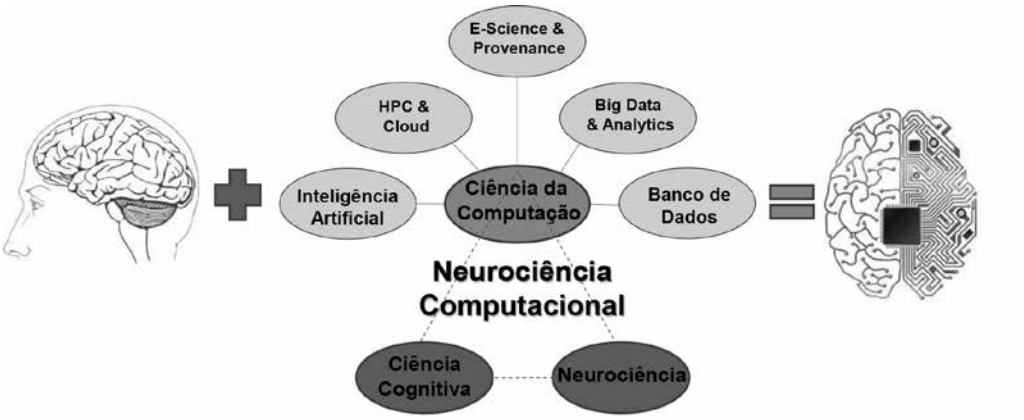


Figura 2. Três pilares na Neurociência Computacional e perspectivas de investigações tecnológicas apoiadas pela Ciência da Computação.

Para alcançarmos um patamar tecnológico compatível com as demandas multifacetadas dos estudos do cérebro no Século XXI, deveremos considerar a estruturação de uma agenda de pesquisa com ênfase em *novos modelos* e incorporação de *técnicas computacionais (intensivas em dados) do E-science* (HEY *et al*, 2009) para o desenvolvimento de aplicações em Neurociência Computacional. Neste caso é possível avançar nas seguintes áreas:

Desenvolvimento de sistemas inteligentes e preditivos baseados em técnicas de Inteligência Artificial (MACÁRIO COSTA *et al*, 2007, 2008, 2009, 2010, 2011, 2013; ZAVALETA *et al*, 2012) capazes de manipular grandes volumes de dados;

Uso ambientes computacionais distribuídos de processamento de alto desempenho para apoiar simulações e experimentos *in silico* baseados em workflows científicos (DEELMAN *et al*, 2009) de simulações de modelos cerebrais (ABBOT, 2013; KUBILIUS, 2014; NIH, 2014; HBP, 2014).

Adoção de técnicas de gestão de grandes volumes de dados biológicos semiestruturados e processamentos típicos de *Big Data* (DAVISON, 2010; BERMAN, 2011; ABBOTT, 2013), os projetos de neurociência computacional tendem mapear modelos cerebrais cada vez maiores, mais complexos e utilizar sensores e dados com diversos formatos (ZHONG *et al*. 2011; ZHONG, 2012);

Incorporação de descritores de proveniência (CRUZ *et al*, 2009) e curadoria de dados e técnicas de gestão de conhecimento para ampliar a reprodutibilidade e a confiabilidade dos estudos em neurociência computacional (CHEN, ZHONG, LIANG, 2012; CICCARESE *et al*, 2013) estes tendem a ser conduzidos por times de pesquisa interdisciplinares e dispersos geográfica e temporalmente (CHEN, ZHONG, 2013);

O conhecimento adquirido na área de neurociência pode ser associado às ferramentas e técnicas computacionais para aperfeiçoar as oportunidades de atuação sobre Transtornos de Aprendizagem.

A dislexia é um Transtorno de Aprendizagem que afeta de 3-7% da população em idade escolar, e destaca de outros transtornos que incluem atrasos graves na leitura, na escrita e ortografia, assim como inversões de símbolos. O destaque se deve a sua natureza única e limitada do déficit fonológico (SHAYWITZ e SHAYWITZ, 1999; MOUSINHO, 2003). A compreensão detalhada das correlações entre variações genéticas, disfunções cerebrais e dificuldades de cognição representa um grande desafio na pesquisa da dislexia (GIRAUD e RAMUS, 2013). Hoje, uma avaliação de um indivíduo disléxico demora em média dois meses até o estabelecimento de um diagnóstico por uma equipe qualificada. E, na maioria das vezes, os escolares encaminhados para o serviço não apresentam dislexia, e sim problemas de aprendizagem de outra ordem. Aqui se defende que através do estabelecimento de novos sistemas de rastreio apoiados na Neurociência Computacional será possível diminuir as filas e agilizar o acesso ao diagnóstico, oferecendo chances de intervenção para mais crianças e jovens em um curto espaço de tempo, de forma mais oportuna, eficiente e socialmente mais justa.

3. CONTRIBUIÇÃO DA TEMÁTICA AOS GRANDES DESAFIOS

A participação no 3º Seminário dos Grandes Desafios em Computação proporcionará a troca de saberes e práticas, bem como a oportunidade de apresentar um projeto de pesquisa investigativo usando técnicas e métodos de inteligência artificial para o tratamento de grandes massas de dados visando o rastreio de pessoas com Transtornos de Aprendizagem.

Investigações direcionadas ao rastreio de crianças e jovens com dislexia contribuirão para estabelecer avaliações relacionadas a identificação e encaminhamento oportuno de pessoas com Transtornos de Aprendizagem na rede ensino, pública ou privada.

Mousinho (2011) coloca que chega ser uma situação rotineira encontrar aluno com dificuldade de leitura. Na maioria das escolas, as turmas são compostas por crianças mais ou menos hábeis em leitura. E, entre as menos hábeis, há ainda aquelas que se destacam. Por mais que se esforcem, a leitura para estas crianças se torna uma atividade laboriosa e até penosa. O esforço colocado nesta ação é tão grande que pode inclusive dificultar o gosto pela leitura.

As crianças com leitura mais comprometida tendem a aumentar a defasagem entre seus pares, o que pode favorecer o aparecimento de consequências indesejáveis como a perda do prazer pela leitura, o baixo rendimento em outras disciplinas que dependem da leitura e o desenvolvimento pela criança de baixa autoestima (MOUSINHO, 2011).

Identificar precocemente crianças com dificuldades de leitura e aprendizagem torna-se prioridade diante da possibilidade de poder eliminar ou minimizar prejuízos escolares e sociais nas mesmas. Dessa maneira, as estratégias computacionais da Neurociência Computacional podem fazer toda a diferença nessa identificação.

Uma das estratégias computacionais da Neurociência Computacional é o uso de redes neurais. As redes neurais são modelos matemáticos que se assemelham às estruturas neurais biológicas e que tem capacidade computacional adquirida por meio de aprendizado e generalização. Consideramos que o maior desafio é reunir todo tipo de informação (variáveis) em bases de dados e desenvolver novos mecanismos para análise e detecção do paciente disléxico.

Nesse sentido esta pesquisa vem se desenvolvendo de forma singular no que diz respeito à detecção de crianças e adolescentes em risco de dislexia. O sistema inteligente desenvolvido por Macário Costa *et al* (2007, 2008, 2009, 2010, 2011, 2013) está ampliando sua base de dados para consolidar mais ainda a veracidade dos algoritmos utilizados para extrair a informação privilegiada dos dados coletados a partir de entrevistas feitas com o responsável da criança que está matriculado na escola, na busca por um padrão desejado para a identificação do indivíduo que apresenta a dificuldade.

No apoio ao diagnóstico dos Transtornos Específicos da Aprendizagem, Macário Costa *et al* (*l.c.*) apresentam uma implementação de uma rede neuronal do tipo *Multi Layer Perceptron* para classificar probabilisticamente pacientes jovens e adultos com dislexia.

4. PLANO PRELIMINAR PARA O DESENVOLVIMENTO

O grupo de pesquisa busca desenvolver soluções computacionais que possam contribuir de forma precoce com o diagnóstico de pessoas com Transtornos de Aprendizagem no nosso país. As soluções computacionais poderão ser aplicadas diretamente pelos professores nas escolas, neste sentido, destacamos a necessidade de buscar parcerias com instituições públicas e privadas, governos e grupos de pesquisa que possam cooperar e oferecer infraestrutura computacional capazes de armazenar e processar de modo distribuído, dados semiestruturados e que também ~~que~~ sejam capazes de oferecerem comprometimento para o envolvimento das suas equipes nas escolas para rastrear os indivíduos com dificuldades de aprendizagem e/ou risco de Dislexia, ou outro tipo de Transtorno de Aprendizagem.

Este projeto é intrinsecamente multidisciplinar e deve ser conduzido por muitas mãos. Ele está dividido em três fases. A primeira, já realizada, se constituiu do delineamento do problema e projeto da ferramenta e das redes neurais; seguido da construção do sistema inteligente e de sua base de dados e da realização dos testes e avaliações dos modelos.

A segunda fase, em progresso, consiste do levantamento das novas demandas necessárias à escalabilidade da solução, tornando-a capaz de incorporar os novos artefatos tecnológicos (discutidos anteriormente) que possam processar grandes volumes de dados em ambientes distribuídos de alto desempenho. Nesta fase também ocorre o mapeamento de novos colaboradores e escolas parceiras. Sem essa combinação de tecnologias, não conseguimos responder à pergunta básica: Como detectar precocemente e com baixo custo operacional e alta eficácia, jovens em risco de dislexia, ou outro tipo de Transtorno de Aprendizagem? Num País como o Brasil, com suas dimensões continentais e suas históricas desigualdades sociais, obter respostas para perguntas dessa natureza pode representar um diferencial estratégico para o futuro do País.

A terceira e última fase é crucial para o sucesso do projeto. Será mais pulverizada e terá caráter operacional. Será voltada para o rastreio massivo e precoce no ambiente escolar das crianças com dificuldades de leitura e de aprendizagem. Esta fase requer análise e tratamento de dados de maneira precisa, confiável e rápida para auxiliar no diagnóstico médico e estabelecer encaminhamento precoce desses indivíduos para os especialistas.

AGRADECIMENTOS

Os autores R.J.M. Costa e R. Mousinho agradecem à equipe do Projeto ELO (Departamento de Fonoaudiologia, Faculdade de Medicina da UFRJ) e do Instituto de Neurologia Delindo Couto da UFRJ e a USAL. S.M.S. Cruz, agradece à FAPERJ e ao MEC/SeSU pelo apoio financeiro às pesquisas.

REFERÊNCIAS

Abbott, A. (2013) “Neuroscience: Solving the brain”, In: Nature 499, pages. 272–274.

Berman JJ, (2009) “Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information” Morgan Kaufmann; 1st Edition.

Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J.G, Goble, C., Clark, T. (2013) “PAV ontology: provenance, authoring and versioning”, Journal of Biomedical Semantics 2013, 4:37.

Chen J. H., Zhong, N. (2013) “Toward the Data-Brain driven systematic brain data analysis”. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43(1), pages. 222-228.

Chen J. H., Zhong, N., Liang, P. P. (2012) “Data-Brain driven systematic human brain data analysis: A case study in numerical inductive reasoning centric investigation” Cognitive Systems Research, Elsevier, vol. 15-16, pages. 17-32.

Cruz, S.M.S, Campos, M. L M., Mattoso, M. (2009) "Towards a Taxonomy of Provenance in Scientific Workflow Management Systems". SERVICES I 2009: 259-266.

Davison, A.P. (2010) "Challenges and solutions in replicability and provenance tracking for simulation projects". BMC Neuroscience 2010, 11(Suppl 1):P76.

Deelman E, Gannon D, Shields M, Taylor I (2009) "Workflows and e-Science: An overview of workflow system features and capabilities", Future Generation Computer Systems 25(5):528–540.

Giraud A.L., Ramus, F. (2013) "Neurogenetics and auditory processing in developmental dyslexia". Curr. Opin. Neurobiol. 23:37–42.

HBP (2014) "Human Brain Project". https://www.humanbrainproject.eu/

Hey, T., Tansley, S., Tolle, K (2009) "The Fourth Paradigm: Data-Intensive Scientific Discovery" Microsoft Press. 1st Edition.

Kubilius, J. (2013) "A framework for streamlining research workflow in neuroscience and psychology". Front. Neuroinform. 7: 52.

Lent, R. (2001) "Cem Bilhões de Neurônios – Conceitos Fundamentais de Neurociência". São Paulo. Ed. Atheneu.

Macário Costa, R. J. *et al*. (2007) "Classificação de pacientes com transtorno de dislexia usando redes neurais artificiais". In: XXX CNMAC, Florianópolis.

Macário Costa, R. J., Mousinho, R.; Vidal, L. A. (2008) "Redes Neuronais: um instrumento no rastreio (screening) de pessoas com risco de transtorno específico de leitura". In: I Congresso Ibro/Larc de Neurociências da América Latina, Caribe e Península Ibérica, 2008, Búzios, RJ.

Macário Costa, R. J., Mousinho, R.; Vidal, L. A. (2009) "Abordagem Computacional no Screening da Dislexia e do TDAH". In: XXXII CNMAC, Cuiabá.

Macário Costa, R. J. *et al*. (2009) "Redes neuronais e transtornos de aprendizagem: rastreio de pessoas com dislexia". In: SBIE. v. 20. p. 1-10. Florianopólis

Macário Costa, R. J., Mousinho, R.; Vidal, L. A. (2009) "Dislexia e Inteligência Computacional: Um sistema para rastrear (Screening) pessoas com sinais de transtorno de leitura". In: 2º. Congresso Internacional de Dislexia, São Paulo.

Macário Costa, R. J. (2011) "Uma Estratégia computacional na detecção da dislexia". Rio de Janeiro: Tese – UFRJ/COPPE.

Macário Costa, R. J. *et al*. (2011) "Abordagem tecnológica para rastreio de pessoas com dislexia". Tecer (Belo Horizonte), v. 4, p. 41-53.

Macário Costa, R. J. *et al.* (2013) "A Computational Approach for Screening Dyslexia". In: CBMS 2013, 2013, Porto. 26th IEEE International Symposium on Computer-Based Medical Systems.

Markram, H. (2013) "Seven Challenges in Neuroscience", In: Functional Neurobiology 28(3) 145-151.

Mousinho, R. (2003) "Desenvolvimento da Leitura, Escrita e seus Transtornos. In: Goldfeld, M. Fundamentos em Fonoaudiologia". Guanabara Koogan. 2ª edição.

Mousinho, R. *et al.* (2011) "Dislexia – Novos temas, novas perspectivas. Wak Editora. Rio de Janeiro.

NIH (2014) "BRAIN 2025: A Scientific Vision", http://www.nih.gov/science/brain/2025/.

Schwartz, E. (1990) "Computational Neuroscience", MIT Press, 1st edition.

Shaywitz, S.E. Shaywitz, B.A. (1999) "Dyslexia In: Swaiman KF, Ashwal S. Pediatric Neurology - Principal e Practice", Connecticut. Ed. Mosby.

Zavaleta, J. *et al.* (2012) "DysDTool: Uma Ferramenta Inteligente para Avaliação e Intervenção no Apoio ao Diagnóstico da Dislexia". In: CSBC- XII Workshop de Informática Médica.

Zhong, N. *et al.* (2011) "Brain Informatics". IEEE Intelligent Systems. September, pages 16-20.

Zhong, N (2012) "Research Issues and Challenges on Brain Informatics Towards Computing & Intelligence in the Big Data Era". pakdd2014.pakdd.org/tutorial3.pdf.

INTEROPERABILIDADE SEMÂNTICA NA CADEIA DE EXPLORAÇÃO DE PETRÓLEO

Mara Abel[1], Luiz Fernando De Ros[2], Joel Carbonera[1],
Sandro Rama Fiorini[1], Alexandre Lorenzatti[1]

Resumo. A cadeia de petróleo consiste em uma sequência de tarefas que inicia com a exploração de áreas potencialmente econômicas por geólogos e geofísicos, produzindo dados que são utilizadas para subsidiar as decisões dos engenheiros de petróleo sobre colocar ou não um determinado reservatório em produção. Cada um desses profissionais possui diferentes arcabouços conceituais influenciados pelo conhecimento específico de cada área do conhecimento em que atuam. Devido a isto, cada profissional analisa a realidade com uma visão que pode ser completamente diferente da visão assumida pelos demais. Durante todo o processo, são gerados dados a respeito de uma mesma realidade compartilhada, mas que são estruturados por conceitualizações distintas. Devido a este cenário, a cadeia de petróleo torna-se uma área rica em desafios relacionados à modelagem conceitual e integração de dados, visando à interoperabilidade semântica. Nós defendemos que as ontologias desempenham um importante papel no desenvolvimento de soluções para integração de dados e interoperabilidade de sistemas na cadeia de exploração e produção de petróleo. Defendemos também que propriedades ontológicas – estudadas em disciplinas como a Ontologia Formal – podem ser utilizadas para caracterizar de modo bem fundamentado os conceitos que são utilizados ao longo da cadeia, viabilizando a interoperabilidade semântica.

1. INTRODUÇÃO

O grupo BDI (Grupo de Bancos de Dados Inteligentes do Instituto de Informática da UFRGS[1]) tem pesquisado por mais de 20 anos aplicações de técnicas da Engenharia de Conhecimento e Engenharia de Ontologias em problemas relacionados à cadeia de petróleo.

A cadeia de petróleo corresponde a um domínio conceitualmente rico, que circunscreve um conjunto variado e complexo de atividades realizadas por especialistas de diferentes áreas, com conhecimentos distintos e, consequentemente, com diferentes visões sobre a realidade. A cadeia como um todo integra as etapas de:

[1]Instituto de Informática e [2]Instituto de Geociências – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{marabel,lfderos,jlcarbonera,srfiorini,alorenzatti}@inf.ufrgs.br

[1]www.inf.ufrgs.br/bdi

Exploração: que se refere a pesquisar e encontrar petróleo em quantidades comerciais;

Produção: na qual o petróleo é extraído do reservatório com rentabilidade, segurança e respeito ao meio ambiente;

Transporte: cujo objetivo é conduzir o petróleo bruto e seus derivados das áreas de produção para as de refino e consumo;

Refino: na qual são obtidos do petróleo os produtos necessários à sociedade;

Comercialização: que busca fornecer os derivados onde eles são necessários, a preços competitivos.

Cada uma destas etapas inclui um conjunto de atividades intensivas em conhecimento, que são realizadas de modo encadeado, de modo que cada atividade depende dos dados gerados pelas atividades precedentes, e cada uma é conduzida por profissionais com conceitualizações distintas acerca da mesma realidade. O suporte às atividades é oferecido por muitas dezenas de aplicações especializadas desenvolvidas por diferentes empresas de software ao longo dos últimos 50 anos. Recentemente, a integração de dados ao longo da cadeia de E&P tem sido considerada como um dos desafios mais importantes a serem atacados pela indústria. As quantias aplicadas no desenvolvimento de soluções de integração igualam a aquisição de novos produtos de software.

A indústria de petróleo, órgãos de governo e a academia têm se unido em associações e consórcios com o objetivo de propor soluções conjuntas para o problema da interoperabilidade e integração de dados. Em termos mundiais, onze destas associações estão conjuntamente representadas pelo *Standards Leadership Council* (SLP)[2], o conselho que une os principais órgãos e associações de regulamentação de padrões da indústria mundial de óleo e gás, com foco na criação de um conjunto de padrões abertos nos mais diferentes segmentos da indústria

No Brasil, a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP [3] – centraliza os esforços para a adoção de padrões e integração de dados para a indústria de petróleo e disponibiliza as informações públicas através do BDEP – Banco de Dados de Exploração e Produção[4].

O Grupo BDI participa das associações pertencentes ao SLP cujo foco é principalmente voltado à interoperabilidade de dados nas atividades da exploração de petróleo. Estas associações são:

[2]www.oilandgasstandards.org
[3]www.anp.gov.br
[4]http://www.bdep.gov.br/

Energistics[5]: que define padrões para troca de informações entre sistemas que apoiam as fases de *upstream* (anteriores à produção propriamente dita);

Professional Petroleum Data Management Association (PPDM)[6]: que define modelos de dados e disseminação de melhores práticas em gestão de dados.

As contribuições do grupo se referem à utilização de diferentes abordagens baseadas em ontologias para modelagem de informações com foco na interoperabilidade dos sistemas que oferecem suporte às etapas de exploração da cadeia de óleo e gás. Ontologia, neste sentido, se refere à teoria que lida e justifica o significado pretendido de um vocabulário formal, isto é, lida com o *compromisso ontológico* deste vocabulário com o mundo (Guarino, 1998). Quando materializado em um artefato, ontologias são descritas como especificações formais e explícitas de conceitualizações compartilhadas (Gruber, 1995). Recentemente, ontologias têm sido aplicadas como uma abordagem para lidar com a heterogeneidade semântica, uma vez que elas permitem explicitar de modo formal e não ambíguo o significado dos conceitos representados nos sistemas.

O Grupo BDI defende a utilização de análise ontológica dos conceitos que são representados nos sistemas, realizada com base em teorias suportadas por disciplinas como a Ontologia Formal, como uma forma de explicitar o significado pretendido pelo modelador ao criar modelos. Ainda, o grupo tem produzido ontologias de domínio bem fundamentadas nas áreas de Petrologia e Estratigrafia, domínios centrais para a modelagem geológica. A análise ontológica dos conceitos representados nos múltiplos sistemas e a relação destes conceitos com objetos geológicos formalmente definidos viabiliza a identificação de conceitos semanticamente relacionados (isto é, que se referem às mesmas entidades da realidade), promovendo a integração de informações.

Estes resultados têm sido recebidos e adotados pelas associações de promoção de padrões acima listadas, e incentivado a disseminação da utilização de ontologias para o desenvolvimento de soluções para os problemas de integração de informação da cadeia de Petróleo, em especial nas atividades de *upstream*.

Na seção 2, discutiremos em mais detalhes os desafios de integração de informações relacionados à cadeia de Petróleo. Na seção 3, apresentaremos o grupo BDI, discutindo as suas principais contribuições em termos de soluções para tais desafios. Na Seção 4, discutimos a interação do grupo com empresas e instituições internacionais. Na Seção 5, apresentamos nossas considerações finais.

[5]www.energistics.org
[6]www.ppdm.org

2. INTEGRAÇÃO DE INFORMAÇÕES NA CADEIA DE PETRÓLEO

Durante a fase de exploração, na qual os reservatórios de petróleo são prospectados e avaliados quanto à economicidade, são geradas grandes quantidades de dados sísmicos, dados de perfuração, dados de sondagens de poços, etc. A sequência de atividades relacionadas ao *upstream* está esquematicamente representada na Figura 1. É a etapa menos estruturada de todo o processo, onde interpretações subjetivas dos dados são utilizadas como entradas para as atividades subsequentes. Esses dados, em geral coletados em procedimentos conduzidas por geólogos, são estruturados por conceitualizações do domínio da Geologia, que são o foco de estudo do Grupo BDI. Na etapa de produção, em que o petróleo é efetivamente extraído dos reservatórios naturais, as atividades são conduzidas majoritariamente por engenheiros; de modo que os dados gerados nesta etapa envolvem conceitualizações da Engenharia. O sucesso das atividades depende do uso eficiente deste grande volume de dados heterogêneos para a construção de modelos que são utilizados para reduzir a incerteza e o risco nos processos de tomada de decisão.

As múltiplas conceitualizações de objetos geológicos presentes na fase de Exploração correspondem a diferentes perspectivas sobre uma mesma realidade subjacente. Por exemplo, modelos de reservatório para simulação de fluxo são desenvolvidos com o intuito de simular a acumulação e o deslocamento do óleo no interior dos volumes de rocha que compõem os reservatórios. Estes modelos são gerados a partir de dados geológicos brutos (descrições de rocha, dados sísmicos, petrofísicos e de análises químicas), como resultado de uma longa cadeia de interpretações sucessivas, realizadas por diferentes profissionais (geofísicos, geólogos e engenheiros de reservatório) com diferentes conceitualizações a respeito da mesma realidade subjacente. Apesar de os dados capturados se referirem a um mesmo objeto geológico na realidade subjacente (no caso uma unidade litológica porosa, com uma estrutura que retém o óleo) os especialistas utilizam aplicações de software com diferentes representações e formatos de codificação para lidar com a mesma informação em diferentes partes do processo de trabalho. Deste modo, neste processo, os dados que são gerados e processados, na verdade, podem se referir a um mesmo conjunto de entidades da realidade, mas que são conceitualizadas de modo distinto em etapas distintas. A heterogeneidade em termos de conceitualização e representação relacionada a estes dados inviabiliza a interoperabilidade entre os sistemas que são utilizados ao longo da modelagem do reservatório.

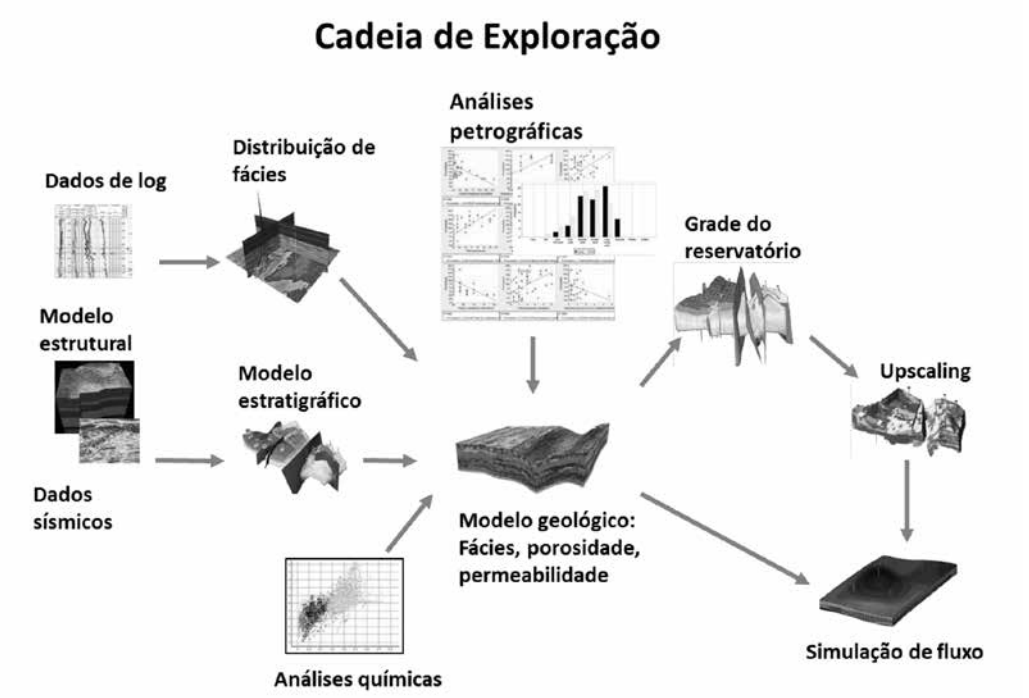


Figura 1: Atividades de geração e modelagem de informações na etapa de exploração. Adaptado de (Perrin, Rainaud *et al.*, 2007)

Interoperabilidade, neste cenário, é a capacidade de oferecer ao usuário o acesso uniforme aos dados ou informações criadas e gerenciadas por diferentes sistemas de informação, permitindo a realização de análises unificadas sobre eles (Wache, Vogele *et al.*, 2001). Isto é, interoperabilidade é a capacidade de sistemas de informação distintos comunicarem entre si, compartilhando dados, informação e conhecimento de forma segura e eficiente.

A interoperabilidade tem sido buscada através da identificação – em cada um dos modelos – dos mesmos objetos geológicos existentes na realidade. A abordagem esbarra no fato de que os modelos capturam modelos mentais parciais dos objetos de estudo que nem sempre permitem o mapeamento entre suas descrições. Esse problema poderia ser melhor tratado através da formalização de uma ontologia geológica de topo, que contivesse uma descrição formal dos principais objetos geológicos que são investigados durante a Exploração. A proposição de ontologias para descrição de objetos

[7]www.cgi-iugs.org
[8]www.geosciml.org

geológicos tem sido uma das abordagens utilizadas pela *Commission for the Management and Application of Geoscience and Information*[7]. Esta comissão é responsável pelo desenvolvimento do *GeoSciML Resource Repository*[8], que busca descrever a definição padrão dos termos geológicos que oferecem suporte aos modelos. Embora o GeoSciMl não tenha em sua proposta a construção de ontologias de domínio para petróleo, seus objetivos têm se apoiado na utilização destas técnicas para a definição da semântica dos conceitos.

Os esforços desenvolvidos nos últimos anos pelos diferentes agentes envolvidos na promoção da interoperabilidade de dados na Exploração de petróleo são limitados pelo estado da arte na modelagem de conhecimento. Alguns dos principais problemas que têm recebido o foco da pesquisa recente em modelagem conceitual são:

• como modelar a multiplicidade de papéis assumidos pelas entidades da realidade e que são considerados como conceitos distintos nos modelos de informação, preservando a identidade dos objetos modelados;

• como modelar a evolução das entidades da realidade, preservando sua identidade através das diferentes fases em que esta entidade atravessa (modelagem de ocorrentes e continuantes);

• quais as primitivas necessárias e suficientes para modelar ontologicamente os diferentes conceitos que as pessoas utilizam para lidar com a complexidade do mundo.

Alguns avanços importantes foram oferecidos pela pesquisa em ontologias formais desenvolvida por Nicholas Guarino da Universidade de Trento, Itália, no estudo das propriedades ontológicas dos conceitos (Guarino, 1992; Gruber, 1993; Guarino, 1995; 1998; Guarino e Welty, 2000) e no desenvolvimento da metodologia Ontoclean (Guarino e Welty, 2002; Guarino e Welty, 2004) para a análise ontológica dos modelos. Paralelamente Barry Smith, da *Saarland University, Saarbrücken* (Neuhaus, Grenon *et al.*, 2004) e Chris Patridge, da Boro Solutions, da Inglaterra, (Partridge, 2005) também propuseram abordagens e primitivas para a modelagem conceitual baseada em ontologias. Com forte influência das propostas de Guarino, Giancarlo Guizzardi, da Universidade do Espírito Santo, propôs a UFO (Unified Foundational Ontology), um arcabouço conceitual completo para a análise ontológica de conceitos, com proposição de primitivas de modelagem em meta-nível (Guizzardi, 2005). Mais recentemente, Guizzardi detalhou as primitivas para tratar aspectos temporais dos conceitos através da UFO-B (Guizzardi, Wagner *et al.*, 2013).

Os trabalhos listados têm influenciado as abordagens ontológicas aplicadas pelo Grupo BDI na construção de ontologias de domínio em Geologia de Petróleo e também na proposição de abordagens para os problemas de interoperabilidade dos sistemas de modelagem geológica para exploração.

3. O GRUPO BDI E SUAS CONTRIBUIÇÕES

O Grupo de Banco de Dados Inteligentes (BDI) foi criado em 1993 no Instituto de Informática da UFRGS, com foco em pesquisa e desenvolvimento de sistemas de informação com capacidades de representação semântica e raciocínio. Em especial, o grupo investiga as possibilidades de construir sistemas de conhecimento capazes de persistir conhecimento em bancos de dados e extrair novas informações com métodos de raciocínio.

O grupo tem uma tradição de interdisciplinaridade, incluindo pesquisadores de Ciência da Computação, Geologia e Comunicação, que procuram fornecer soluções computacionais aplicadas a problemas reais, sendo internacionalmente reconhecido por investigar a aplicação de técnicas da Engenharia de Conhecimento e Engenharia de Ontologias em problemas relacionados à cadeia de Petróleo (Abel, Goldberg *et al.*, 2012). A modelagem conceitual do domínio da Geologia para integração de informações constitui uma das principais linhas de investigação do grupo BDI. Nos projetos realizados pelo grupo nesta linha, foca-se em problemas típicos da etapa de Exploração na cadeia de Petróleo, em especial na modelagem conceitual dos objetos geológicos e seu mapeamento ao longo da cadeia (Perrin, Rainaud *et al.*, 2012).

Uma das principais características do grupo é o esforço que realiza no sentido de aproximar pesquisa acadêmica e indústria, transferindo tecnologia para a sociedade por meio de parcerias com empresas e institutos de pesquisa. Diversos projetos aplicados em parceria com a indústria de diversos segmentos foram desenvolvidos nas últimas duas décadas, porém os resultados mais notáveis foram obtidos em aplicações da indústria do petróleo. Um de seus principais resultados, o Projeto PetroGrapher (Abel, Silva *et al.*, 2004), gerou o sistema Petroledge®[9], comercializado pela empresa Endeeper[10], gerada como um spin-off do Grupo BDI. O Petroledge® é utilizado em apoio à avaliação de qualidade de reservatórios de petróleo (De Ros, Goldberg *et al.*, 2007; Goldberg, Abel *et al.*, 2008; De Ros, Abel *et al.*, 2009) em laboratórios, universidades e empresas internacionais, incluindo a Petrobras.

Os projetos de pesquisa produziram ontologias largamente reconhecidas. A ontologia de petrografia de rochas siliciclásticas proposta em (Abel, 2001) permitiu a criação dos sistemas Petroledge, PetroQuery e RockViewer. Mais recentemente, o grupo desenvolveu as primitivas visuais e a estrutura conceitual necessárias para a criação da ontologia para descrição de fácies sedimentares em Estratigrafia (Lorenzatti, 2010). A ontologia esboçada por Lorenzatti foi posteriormente expandida e permitiu o desenvolvimento da aplicação Strataledge [11], para a descrição de testemunhos de rochas utilizando dispo-

[9]www.endeeper.com/product/petroledge
[10]www.endeeper.com
[11]www.endeeper.com/product/strataledge

sitivos móveis (Abel, Lorenzatti *et al.*, 2012), oferecendo suporte para a integração de dados de descrição petrográfica de rochas-reservatório com dados estratigráficos capturados de testemunhos de poços de exploração de petróleo (Ros, Goldberg *et al.*, 2009).

As ontologias possibilitam a construção de sistemas de conhecimento para apoio das tarefas em Exploração, ampliando as possibilidades do desenvolvimento de raciocínio automático para extração de interpretações geológicas. O trabalho de Mastella, por exemplo, permite a interpretação automática dos eventos diagenéticos que afetam as condições de porosidade e permeabilidade em reservatórios siliciclásticos (Mastella, 2004; Mastella, Abel *et al.*, 2005; Mastella, Abel *et al.*, 2007). Santin utilizou descrições baseadas em ontologias para a detecção do grau de compactação dos reservatórios (Santin, 2007; Santin, Abel *et al.*, 2009). Dutra desenvolveu um trabalho no Curso de Geologia para identificar os parâmetros descritos com sistema Petroledge que permitem reconhecer petrofácies relacionadas à qualidade de reservatórios (Goldberg, Abel *et al.*, 2008; Soares, 2008; Soares, Goldberg *et al.*, 2008). Mais recentemente, utilizando abordagens de ontologias formais, Carbonera busca extrair a interpretação de sequências de eventos de transporte e deposição para análise estratigráfica em bacias sedimentares (Carbonera, Abel *et al.*, 2011; Carbonera, 2012; Carbonera, Abel *et al.*, 2013).

Outra linha de investigação do Grupo busca a integração de dados geofísicos extraídos de poços de exploração com os modelos ontológicos. Fiorini estudou a extração de objetos geológicos a partir de registros geofísicos de poços com base nos métodos de identificação visual aplicados por geólogos (Fiorini, 2009; Fiorini, Abel *et al.*, 2010; Fiorini, Abel *et al.*, 2011; 2012). Utilizando descrições de testemunhos de poços baseadas em ontologias, Garcia (Garcia, Carbonera *et al.*, 2013) investiga formas de extrair correlações automáticas entre poços e também as correlações das descrições de rocha com os perfis geofísicos do poço.

Estes estudos tornaram clara a vantagem de utilizar modelos conceituais bem fundamentados para a integração de informações geológicas e geofísicas. Werlang vem investigando a integração de padrões de troca de informações na indústria de petróleo através da identificação dos objetos geológicos mencionados nos padrões (Werlang, Abel *et al.*, 2013).

O processo de integração de informações inicia-se ao determinar se as fontes de dados contêm informações semanticamente relacionadas, isto é, informação sobre entidades que são idênticas (ou similares) na realidade. Para lidar com este problema, o grupo BDI tem utilizado ontologias de fundamentação como uma abordagem basilar no desenvolvimento de soluções.

No grupo BDI, nós defendemos a visão de que a explicitação do significado dos conceitos representados nos modelos é uma condição necessária para promover a interope-

rabilidade entre os modelos desenvolvidos e manipulados por diferentes profissionais, em diferentes sistemas de informação(Abel, Mastella *et al.*, 2012). Além disso, defendemos também o uso de ontologias de domínio desenvolvidas com o suporte teórico de disciplinas, que incluem a Ontologia Formal e as Ciências Cognitivas, como ferramentas adequadas para explicitar a semântica dos conceitos incorporados nos modelos utilizados pela indústria petrolífera(Abel, Perrin *et al.*, 2014).

4. INTERAÇÃO COM EMPRESAS E INSTITUIÇÕES INTERNACIONAIS.

Os trabalhos de pesquisa do grupo BDI são desenvolvidos em parceria com o meio produtivo. Em especial, o grupo mantém relacionamento de longo prazo com a empresa Petrobras, que nos últimos 7 anos tem garantido suporte parcial aos projetos do grupo através de projetos, e do teste e utilização dos sistemas resultantes da pesquisa. O Grupo mantém cooperação também com a microempresa Endeeper, criada como um *spin off* dos produtos de pesquisa do Grupo.

Outras interações importantes do grupo BDI acontecem desde 2003 com o professor Michel Perrin, da École *de Mines de Paris*, e com o Geofísico Dr. Jean-François Rainaud, do Instituto Francês de Petróleo, França; ambas renomadas instituições mundiais de pesquisa em Geologia Estrutural e modelagem de bacias. O objetivo da interação é a definição conjunta de uma ontologia bem fundamentada dos termos gerais em Geologia que permita a construção de sistemas e métodos formais de captura e integração de dados. Um grande número de publicações significativas foi gerado desta parceria.

5. GRANDES DESAFIOS E SUAS SOLUÇÕES

Em resumo, o grupo BDI reconhece pelo menos três grandes desafios para a área de Computação aplicada a cadeia de Petróleo:

Interoperabilidade semântica entre sistemas. É essencial que a tomada de decisões estratégicas na cadeia de Petróleo seja baseada em informações com alto grau de qualidade e confiabilidade. Para isso, é necessário que a informação flua com facilidade pelos diversos sistemas que produzem e processam informações ao longo da cadeia. *Diferentes sistemas e seus projetistas e desenvolvedores devem ter definições claras e completas do significado do vocabulário e estruturas das informações que fluem ao longo de* toda *a cadeia de petróleo*. Mais importante, estes termos e estruturas devem refletir o conhecimento de domínio dos especialistas que consomem e produzem a informação, levando em conta, por exemplo, diferentes visões que possam existir sobre a mesma informação. O grupo BDI defende que *o cerne da solução para o problema de integrar o significado da informação em diversos sistemas seja o desenvolvimento e uso de modelos conceituais bem fundamentados*. Estes devem ser construídos com base em uma me-

todologia que evidencie a natureza da informação, sua evolução e sua utilização. Em particular, modelos conceituais na cadeia de Petróleo devem permitir a representação de diferentes visões sobre o mesmo objeto da realidade subjacente, bem como a representação dos diferentes papeis que a própria informação assume ao longo do processo. Do ponto de vista do desenvolvimento de sistemas, a solução para a interoperabilidade semântica deve incluir metodologias e ferramentas que balizem o emprego de modelos conceituais na implementação de sistemas de informação.

Geração e rastreamento de informação e dados. Informações geradas ao longo da cadeia são geralmente frutos da interpretação de dados brutos coletados em campo. A fim de que a qualidade da informação gerada ao longo da cadeia seja mantida, *deve ser possível rastrear e manter uma ligação explicita entre a informação gerada e os dados que embasaram a sua geração*. Esta ligação garante que o processo de geração das informações que baseiam o processo de tomada de decisão possa ser avaliado e, eventualmente, refeito. Além disso, esta ligação deve facilitar o processo de revisão das informações geradas à medida que novos dados são gerados ou reamostrados. O grupo BDI acredita que esta solução requer o desenvolvimento de novos arcabouços de representação de informação. Estes devem lidar com a natureza simbólica da informação e a natureza continua dos dados, bem como os aspectos dinâmicos de ambos. A formalização desta ligação também abrirá portas para o desenvolvimento de sistemas que *gerem* informação (semi)automaticamente, através da aplicação de técnicas de inteligência artificial e processamento numérico de grandes volumes de dados.

Formação de recursos humanos. A gestão da informação na cadeia de petróleo (e outras áreas) não é trivial, uma vez que se trata de domínio altamente especializado. Desenvolvedores e engenheiros de software que atuam na área devem estar preparados para entender a complexidade da natureza da informação com que lidam. No entanto, a formação destes profissionais não os mune de conhecimento avançado de como traduzir a realidade de domínios complexos para modelos conceituais avançados. Assim, um grande desafio é a *formação de profissionais aptos a gerar, manter e consumir modelos conceituais avançados*. O grupo BDI defende que a solução para este problema seja a inclusão de conceitos de modelagem conceitual avançada em cursos de Engenharia de Software e modelagem de sistemas no país. Esta formação deve permitir novos profissionais tenham mais subsídios teóricos e metodológicos para abstrair a realidade onde atuam em modelos formais semanticamente ricos.

Dado o estado da arte e a experiência adquirida até aqui, o grupo BDI propõe algumas possíveis ações para atacar estes desafios. Estas ações são de grande escopo e em si incluem seus próprios desafios de implementação:

Desenvolvimento de padrões baseados em ontologias. Embora padrões para troca de informações na cadeia de Petróleo existam, a pouca atenção dada à natureza on-

tológica das entidades definidas nesses padrões ao longo do seu desenvolvimento faz com que apresentem ambiguidades definicionais e estruturais. O grupo BDI propõe o desenvolvimento de *padrões baseados em ontologias formais* que definam as principais entidades conceituais na cadeia de Petróleo de forma clara e não ambígua. Eles servirão como base para o desenvolvimento e integração de sistemas na cadeia. Os padrões também devem ser especificados com base nas metodologias de criação de ontologias formais, utilizando conhecimento especialista, e com participação de entidades da academia, governo e indústria.

Evolução de formalismos de representação de informação e conhecimento. Os atuais formalismos de representação de conhecimento são adequados para representação de informações simbólicas estáticas em domínios. Porém, a especificação completa de domínios como a cadeia de Petróleo necessita de formalismos de representação capazes de lidar com informações dinâmicas e dados não-simbólicos. Em particular, a geração e rastreamento de informações e dados ao longo da cadeia requerem arcabouços de representação com estas características. Estes arcabouços devem quebrar com o paradigma estritamente lógico dos arcabouços atuais, incluindo, por exemplo, paradigmas geométricos e associacionistas.

6. CONCLUSÃO

Antes de projetar um aplicativo de software, um modelo conceitual é definido de modo a simplificar a realidade geológica e ajudar a compreendê-lo. Esse modelo preserva a intenção do modelador (geólogo) na concepção da aplicação. No entanto, os modelos não representam a realidade em si, mas sim a conceituação de um modelador sobre esta realidade. Ontologias estão sendo desenvolvidas e utilizadas no domínio da Geologia, a fim de tornar explícita a semântica de modelos de reservatórios e apoiar a integração dos dados gerados nas diversas etapas da cadeia de exploração. No entanto, a definição de conceitos geológicos tem se mostrado uma questão complexa, uma vez que o mesmo vocabulário pode descrever diferentes porções da realidade quando aplicado por um geofísico ou um geólogo, quando desenvolvem modelos geológicos. O Grupo BDI tem aplicado ontologias de fundamentação para ajudar na tomada de decisões ontológicas para a integração de modelos na exploração de petróleo. A metodologia ajuda a validar taxonomias e lida com questões de identidade dos conceitos e suas propriedades ontológicas, a fim de determinar quais objetos mantêm a identidade ao longo de toda a cadeia de modelagem e podem ser utilizados para fins de integração.

6. REFERÊNCIAS

Abel, M. Estudo da perícia em petrografia sedimentar e sua importância para a engenharia de conhecimento. (Tese de Doutorado). Programa de Pós-graduação em Computação, UFRGS, Porto Alegre, 2001. 239 p.

Abel, M., K. Goldberg e L. F. D. Ros. Ontology-based rock description and interpretation. In: M. Perrin e J.-F. Rainaud (Ed.). Knowledge Driven Earth Modelling. Paris: Editions Technip, v.1, 2013. Ontology-based rock description and interpretation, p.268-271

Abel, M., A. Lorenzatti, L. F. D. Ros, O. P. D. Silva, A. Bernardes, K. Goldberg e C. Scherer. Lithologic Logs in the Tablet through Ontology-Based Facies Description. AAPG Annual Convention & Exhibition Long Beach, CA: AAPG 2012.

Abel, M., L. Mastella, M. Perrin e M. Tonnat. Ontologies and their use for geological knowledge formalization. In: M. Perrin e J.-F. Rainaud (Ed.). Knowledge Driven Earth Modelling. Paris: Editions Technip, 2013. Ontologies and their use for geological knowledge formalization, p.189-205

Abel, M., M. Perrin e J. L. Carbonera. Ontological analysis for information integration in Geomodeling. Earth Science Informatics 2014.

Abel, M., L. a. L. Silva, L. F. De Ros, L. S. Mastella, J. A. Campbell e T. Novello. PetroGrapher: Managing petrographic data and knowledge using an intelligent database application. Expert Systems with Applications, v.26, n.1 SPECISS, p.9-18. 2004.

Carbonera, J. Raciocínio sobre conhecimento visual. (Dissertation). Programa de Pos-Graduação em Computação, UFRGS, Porto Alegre, 2012.

Carbonera, J., M. Abel, C. M. S. Scherer e A. K. Bernardes. Reasoning over visual knowledge. Joint IV Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies (ONTOBRAS/MOST). R. Vieira, G. Guizzardi*, et al*. Gramado: CEUR-WS. 776**:** 49-60 p. 2011.

Carbonera, J. L., M. Abel, C. M. S. Scherer e A. Bernardes. Visual Interpretation of Events in Petroleum Geology. IEEE International Conference on Tools with Artificial Intelligence (ICTAI). Washington DC: IEEE 2013.

De Ros, L. F., M. Abel, K. Goldberg e Anonymous. Tackling complex reservoirs through systematic petrographic characterization. AAPG Annual Convention & Exhibition 2009.

De Ros, L. F., K. Goldberg, M. Abel, F. I. Victoretti, L. S. Mastella e E. E. Castro. Advanced Acquisition and Management of Petrographic Information from Reservoir Rocks Using the PETROLEDGE ® System. AAPG Annual Convention & Exhibition Long Beach, CA. April 1-4 2007.

Fiorini, S. R. S-Chart: Um Arcabouço para Interpretação Visual de Gráficos. (Dissertação de Mestrado). Programa de Pós-graduação da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

Fiorini, S. R., M. Abel e C. M. S. Scherer. A Symbol Grounding Model for Semantic Interpretation of 2-D Line Charts. VORTE 2010 - Joint 5th International Workshop on Vocabularies, Ontologies and Rules for The Enterprise (VORTE) - International Workshop on Metamodels, Ontologies and Semantic Technologies (MOST). Vitoria: IEEE 2010.

Fiorini, S. R., M. Abel e C. M. S. Scherer. Semantic image interpretation of gamma ray profiles in petroleum exploration Expert Systems with Applications, v.38, n.4, April 2011 p.3724-3734 2011.

Fiorini, S. R., M. Abel e C. M. S. Scherer. An approach for grounding ontologies in raw data using foundational ontology. Information Systems, December 3, 2012.

Garcia, L. F., J. L. Carbonera e M. Abel. Ontologias Aplicadas ao Problema de Correlação Litológica no Domíinio da Geologia do Petróleo. Joint VI Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies (ONTOBRAS/MOST). Belo Horizonte: CEUR-WS. 1041**:** 203-208 p. 2013.

Goldberg, K., M. Abel, J. A. Daudt, L. F. De Ros, C. Scherer e Anonymous. Reservoir petrofacies of the Echinocyamus Formation (Talara Basin, Peru); an approach for high -resolution reservoir characterization. AAPG Annual Convention & Exhibition,, January 1, 2008.

Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Deventer, The Netherlands: Kluwer Academic Publishers, 1993. p.

Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human-Computer Studies, n.43, p.907-928. 1995.

Guarino, N. Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge bases. Data & Knowledge Engineering, v.8, n.3, p.249-261. 1992.

Guarino, N. Formal ontology, conceptual analysis and knowledge representation. International Journal Human-Computer Studies, v.43, n.2/3, p.625-640. 1995.

Guarino, N. Formal Ontology in Information Systems Formal Ontology in Information Systems, FOIS'98. Trento, Italy**:** 6-8 June 1998.

Guarino, N. e C. Welty. A Formal Ontology of Properties. The ECAI-2000 Workshop on Applications of Ontologies and Problem-Solving Methods. Berlin, Germany: IOS Press 2000.

Guarino, N. e C. Welty. Evaluating ontological Decisions with Ontoclean. Communications of the ACM v.45, n.2, February 2002, p.61 – 65.

Guarino, N. e C. A. Welty. An overview of OntoClean. In: S. Staab e R. Studer (Ed.). Handbook of Ontologies. Berlin: Springer 2004. An overview of OntoClean, p.151-171. (International Handbook on Information Systems)

Guizzardi, G. Ontological Foundations for Structural Conceptual Models. Enschede, The Netherlands: Universal Press, v.05-74. 2005. 410 p. (CTIT PhD Thesis Series)

Guizzardi, G., G. Wagner, R. D. A. Falbo, R. S. S. Guizzardi e J. P. A. Almeida. Towards Ontological Foundations for the Conceptual Modeling of Events. In: W. Ng, V. C. Storey, *et al* (Ed.). Conceptual Modeling - ER 2013. Berlin Heidelberg: Springer-Verlag 2013. Towards Ontological Foundations for the Conceptual Modeling of Events, p.327-341. (Lecture Notes in Computer Science LNCS)

Lorenzatti, A. Ontologia para Domínios Imagísticos: Combinando Primitivas Textuais e Pictóricas. (Dissertação de Mestrado). Programa de Pós-graduação da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

Mastella, L. Um modelo de conhecimento baseado em eventos para aquisição e representação de seqüências temporais em Petrografia Sedimentar. (Dissertação de Mestrado). Programa de Pós-graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

Mastella, L. S., M. Abel, L. C. Lamb e L. F. De Ros. Cognitive Modelling of Event Ordering Reasoning in Imagistic Domains. International Joint Conference in Artificial Intelligence. Edimburgh. 30 July to 5 August 2005.

Mastella, L. S., M. Abel, L. F. D. Ros, M. Perrin e J.-F. Rainaud. Event Ordering Reasoning Ontology applied to Petrology and Geological Modelling. In: O. Castillo, P. Melin, *et al* (Ed.). Theoretical /Advances and Applications of Fuzzy Logic and Soft Computing.: Springer-Verlag, v.42, 2007. Event Ordering Reasoning Ontology applied to Petrology and Geological Modelling, p.465-475. (Advances in Soft Computing)

Neuhaus, F., P. Grenon e B. Smith. A Formal Theory of Substances, Qualities, and Universals. 3rd International Conference on Formal Ontology in Information Systems, FOIS. N. Guarino. Turim 2004.

Partridge, C. Business Objects: Re-engineering for re-use. Huntingdon: The BORO Centre. 2005.

Perrin, M., J.-F. Rainaud e M. Abel. Earth models as subsurface representations. In: M. Perrin e J.-F. Rainaud (Ed.). Knowledge Driven Earth Modelling. Paris: Editions Technip, 2012 Earth models as subsurface representations, p.3-24

Perrin, M., J.-F. Rainaud, L. S. Mastella e M. Abel. Knowledge related challenges for efficient data fusion. Data Fusion: Combining Geological, Geophysical and Engineering Data,SEG/AAPG/SPE Joint Workshop. Vancouver: AAPG 2007.

Ros, L. F., K. Goldberg, C. M. S. Scherer, J. Kuchle, E. S. E. Castro e M. Abel. Integrated Petrographic, Stratigraphic and Statistical Analysis of Complex Albian Reservoirs in the Espírito Santo Basin, Eastern Brazil. AAPG Annual Convention & Exhibition M. Mello. Rio de Janeiro, 2009.

Santin, C. E. Construtos Ontológicos para Representação Simbólica de Conhecimento Visual. (Dissertação de Mestrado). Programa de Pós-Graduação em Computação, UFRGS, Porto Alegre, 2007. 88 p.

Santin, C. E., M. Abel, K. Goldberg, L. F. De Ros e Anonymous. Automatic detection of the degree of compaction in reservoir rocks based from visual knowledge. AAPG Annual Convention & Exhibition Denver: AAPG. January 1, 2009. p.

Soares, R. D. Formalização dos Parâmetros Petrográficos de Impacto sobre a Porosidade e Permeabilidade das Rochas-Reservatório Clásticas e sua Aplicação na Definição de Petrofácies. (Trabalho de Conclusao). Instituto de Geociências, UFRGS, Porto Alegre, 2008.

Soares, R. D., K. Goldberg, M. Abel, L. F. De Ros e Anonymous. Petrofacies de reservatorio; uma ferramenta para a caracterizacao e modelagem otimizada de reservatorios de hidrocarbonetos. Anais do Congresso - Sociedade Brasileira de Geologia, v.44, January 1, 2008.

Wache, H., T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann e S. Hubner. Ontology-Based Integration of Information - A Survey of Existing Approaches. IJCAI-01 Workshop on Ontologies and Information Sharing. A. Gómez-Pérez, M. Gruninger*, et al*. Seattle, USA, : CEURS-WS. 47**:** 108-118 p. 2001.

Werlang, R., M. Abel, M. Perrin, J. L. Carbonera e S. R. Fiorini. Ontological foundations for petroleum application modeling. The Petroleum Network Education Conference,. Houston 2013.

11.1 ARTIGOS

ALAVANCA ESTRATÉGICA DE SOFTWARE E SERVIÇOS DE TECNOLOGIA DA INFORMAÇÃO E DA COMUNICAÇÃO

R Dahab[1]; Michele Nogueira[2]

Resumo

Este documento expressa o interesse dos seus autores de participarem do Terceiro Seminário dos Grandes Desafios em Computação, na sua Fase 2, no desafio de *Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos* e domínio de *Defesa Cibernética*. Queremos deixar claro, porém, que a iniciativa deste documento não limita-se aos seus autores, que agem, sim, como porta-vozes da **Comissão Especial em Segurança da Informação e de Sistemas Computacionais (CESeg)** da Sociedade Brasileira de Computação (SBC). A CESeg tem procurado, desde a sua criação, promover ações integradas dos seus membros e esta chamada nos parece uma oportunidade ímpar para consolidar essa integração e efetivamente agir como um reduto de competência fortalecendo suas conexões com profissionais da área de segurança e os corpos governamentais. Nos apoiamos, naturalmente, nos pilares da pesquisa científica, inovação tecnológica e formação de recursos humanos na área de Segurança Computacional.

1 CONTEXTUALIZAÇÃO E TÓPICOS NORTEADORES

A área de Segurança Computacional, também chamada de Segurança Cibernética entre outros nomes, dispensa argumentos que motivem sua importância técnico-científica, estratégica ou social. Praticamente todas as áreas de pesquisa científica contemplam hoje algum aspecto da Segurança Computacional; bem como todas as áreas de convivência social necessitam de salvaguardas para proteger dados do mau-uso ou de exposição indevida. Do ponto de vista estratégico, a segurança nacional e econômica do Brasil dependem do funcionamento confiável de toda a sua infraestrutura crítica.

[1]Instituto de Computac¸a˜o, UNICAMP - Email: rdahab@ic.unicamp.br
[2]Departamento de Informa´tica, UFPR - Email: michele@inf.ufpr.br

A grande área de Segurança Cibernética começa hoje a amalgamar-se com outras áreas correlatas, formando um quadro de interdisciplinaridade muito rico. O próprio texto desta chamada menciona a busca por "sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos". Sem o provimento de Segurança, no entanto, nenhuma dessas características pode ser atingida adequadamente. Sem a intenção de sermos exaustivos, acreditamos que os seguintes tópicos ocuparão lugar de destaque na pesquisa científica-tecnológica nos anos vindouros:

• **Desenvolvimento de software seguro.** Este tópico é um tanto complexo e merece maior atenção. Software de segurança (em geral criptográfico) deve ser escrito de forma sistemática, com alguma metodologia a guiar o desenvolvimento, e evitando armadilhas conhecidas. Por outro lado, software geral também pode ter um sem-número de vulnerabilidades e raras são as vezes que a indústria de software desenvolve esforços consideráveis para saná-las. Isso se dá pelo simples fato de que a pressão por prazos quase sempre relega requisitos de segurança, não-funcionais, a um segundo plano. Há que se desenvolver importantes esforços nos cursos de Engenharia e Bacharelado de Computação no sentido de incutir nos alunos, desde os seus primeiros passos como programadores, conceitos de programação segura. Além disso, metodologias para o desenvolvimento de software crítico (no sentido da Segurança, não dos requisitos funcionais) estão num estágio muito precário de efetividade. Sem um esforço de sistematização de tais metodologias, estamos fadados a repetir erros e remendar vulnerabilidades.

• **Análise arquitetural de Segurança.** Metodologias balizam o desenvolvimento de software mas nem tudo sai como o planejado, como todo Engenheiro de Software pode atestar: má especificação, incompatibilidades com plataformas, mudanças de curso mal-executadas estão entre algumas razões. Assim, ferramentas para análise arquitetural de Segurança baseadas em métodos formais seriam de grande valia no processo de desenvolvimento. Modelagem de vulnerabilidades, entretanto, é uma tarefa difícil e multifacetada. Novamente, a interação do hardware com o software está no centro das discussões. Entender a natureza da vulnerabilidade, se de software ou hardware, pode ser assunto de discussões infindáveis e não há consenso sobre como formalizá-la.

• **Software de segurança com eficiência energética.** A multiplicação dos dispositivos portáteis e a iminência da Internet das Coisas, colocou a área de eficiência energética dos softwares de segurança (quase sempre um *overhead* ao propósito final do software), no topo da lista das preocupações dos desenvolvedores de software criptográfico e de segurança em geral. Há intensa pesquisa já em curso na comunidade para alcançar eficiência energética sem enfraquecer requisitos de segurança, sem descuidar-se de vazamentos por canais laterais (side-channel).

• **Segurança baseada em hardware.** Outro assunto que vem ganhando relevância, e isso tende a aumentar, é o uso do hardware como último refúgio depositário de

informações críticas, chaves criptográficas, etc. Ideal na teoria, várias formas de invasão vem atormetando a vida da comunidade: canais laterais são um deles; outros são ataques às redes internas de SoCs (systems-on-chip), as chamadas NoCs, que roteiam pacotes entre componentes do circuito integrado, como os múltiplos núcleos nas modernas arquiteturas. Além disso, a interação entre sistema operacional e hardware é outra fonte de preocupação, levando a fabricantes, como a Intel, a propor processadores com segurança embutida, acessível diretamente pelas aplicações (arquitetura SGX). Este tópico intersecta-se com o de análise arquitetural de segurança, já que a descoberta de uma vulnerabilidade após um chip estar no mercado pode ser um desastre para o produtor.

• **Resiliência a intrusões.** Após anos de esforços na detecção de malware de forma a prevenir sua ação, a comunidade parece estar convencida de que a essa luta de "gato e rato" está fadada ao fracasso, e que o melhor é centrar esforços no desenvolvimento de sistemas robustos, que suportem intrusões sem perder sua funcionalidade. Esta área vai aproximar ainda mais as comunidades de dependabilidade (tomada no sentido mais restrito, de tolerância a faltas) da de Segurança, algo que já vem acontecendo, inclusive, no âmbito da CESeg.

Na Seção 2, enumeramos ações estratégicas para o desenvolvimento dos 3 pilares: **pesquisa científica**, **inovação tecnológica** e **formação de recursos humanos**. A Seção 3 motiva a necessidade vital de que o meio acadêmico aumente sua interação com os profissionais e indústrias de Tecnologia da Informação e da Comunicação (TIC) na área de Segurança no Brasil, bem como o governo. A Seção 4 dá um panorama gráfico dos grupos de pesquisa em Segurança no Brasil. A Seção 5 descreve iniciativas em outros países do mundo. A Seção 6 brevemente conclui o documento.

2 AÇÕES

Ações devem ocorrer nos mais diversos níveis e das mais diversas formas. O agentes dessas ações deveriam ser definidos dentro de uma política global de governo, sem relevar iniciativas isoladas que estejam em consonância com essa política. Enumeramos abaixo algumas delas:

• Elaborar, fomentar e implementar programas de Pós-graduação/Capacitação em Segurança. Já há iniciativas da CAPES na área de Ciências Forenses mas um escopo mais largo deveria ser considerado.

• Propostas para inserção de disciplinas da área nos currículos de Computação e Engenharias, especialmente a de Programação Segura, conforme discutimos acima, que, além do mais, é um vetor excelente para conscientização em Segurança

• Elaboração frequente de documentos de análise do contexto de Segurança da Informação no Brasil e no mundo, e suas atualizações, com recomendações para diretrizes de governo e política industrial;

• Quanto a eventos:

• O SBSeg existe há 14 anos mas deveria consolidar sua atuação anual, trazendo com mais frequência setores do governo, da indústria e profissionais. Também deveria gozar de financiamento especial, dado o interesse estratégico da área.

• Eventos correlatos do CGI.br, RNP e SBT deveriam receber uma linha de financiamento mais específica à área de Segurança.

• Trazer um ou dois grandes eventos internacionais nos próximos 5 anos, na área de Segurança, sob os auspícios da CESeg.

• Aproximação com instituições correlatas pelo mundo traria uma bem-vinda oxigenação à área no Brasil.

• Aproximação com outros grupos nacionais, e mesmo com outros grupos dentro de CEs da SBC com interesse em Segurança.

• Aproximação com a indústria e profissionais da area. Esta ação é muito importante e tem sido executada com alguma sucesso com a indústria, mas não com profissionais. A aproximação com profissionais é vital para conhecer o dia-a-dia dos que combatem a batalha diária contra malfeitorias, vazamentos involuntários, etc.

• Fomentar a criação do Portal Nacional de Segurança Computacional. A CESeg se propõe a executar esta tarefa.

• Fomentar a participação em competições nacionais e internacionais da area.

• Lançamento de uma publicação científica de qualidade na area. Precisa estar inserida em ações globais para definição de escopo, política de publicação, etc.

3 A INTERAÇÃO COM A INDÚSTRIA, PROFISSIONAIS DE SEGURANÇA E GOVERNO NO BRASIL

A interação com a indústria brasileira tem sido de relativo sucesso: o Intel Labs, ramo de pesquisa da Intel Semicondutores, firmou contrato no Brasil para financiamento de 6 projetos na área de Segurança com Eficiência Energética; o valor total dos contratos está na ordem de 1 milhão de dólares. Como contrapartida, o CNPq aprovou várias bolas para que esses projetos tivessem continuidade. A Intel também está financiando a produção de material didático na área de segurança em duas instituições

brasileiras. Outras empresas como Samsung, LG, Lenovo e outras, usando benefícios da lei de informática, tem trazido recursos valiosos para os grupos de pesquisa na área no Brasil. Acordos com o setor público também têm surgido, com o Banco do Brasil, Departamento de Polícia Federal, Receita Federal, entre outros. Especificamente com órgãoes do governo, a SBC, por meio da CESeg, continua atuando no Comitê Gestor da Infra-estrutura de Chaves Públicas Brasileira. Pesquisadores da comunidade também produziram o primeiro hardware de segurança máxima com certificação internacional, que equipa a Autoridade Certificadora Raiz da ICP-Brasil. Há iniciativas também que surgiram dentro do governo mas com pouca interação com a SBC, como por exemplo a RENASIC, que congrega grande número de profissionais e acadêmicos. Essa interação precisa ser incrementada.

4 PANORAMA DOS GRUPOS NO BRASIL

O mapa abaixo dá uma ideia da distribuição dos grupos de pesquisa no Brasil.

5 POSICIONAMENTO DO BRASIL NO CENÁRIO MUNDIAL

Um dos primeiros países a reconhecer a segurança cibernética como uma questão estratégica nacional foram os Estados Unidos da América. Em 2003, esse país publicou a

Estratégia Nacional para Segurança Cyberspace. Era uma parte da estratégia nacional global para a Segurança Interna, que foi desenvolvido em resposta aos ataques terroristas em 11 de Setembro de 2001.

Desenvolvido por razões semelhantes, os planos e as estratégias com foco limitado de ação começaram a surgir em toda a Europa nos anos seguintes. Em 2005, a Alemanha aprovou o "Plano Nacional de Proteção de Informações de Infra-estrutura (NPSI)". No ano seguinte, a Suécia desenvolveu uma "Estratégia para melhorar a segurança na Internet na Suécia". Após a grave ciberataque na Estônia, em 2007, o país foi o primeiro Estado-Membro da União Européia (UE) a publicar uma ampla estratégia nacional de segurança cibernética em 2008. Desde então, um trabalho considerável tem sido feito nesta área, em nível nacional e nos últimos quatro anos, os dez Estados-Membros da UE têm publicado uma estratégia nacional de segurança cibernética. Estes são resumidos abaixo:

• Estônia (2008): Estônia enfatiza a necessidade de um ciberespaço seguro em geral, e concentra-se em sistemas de informação. As medidas recomendadas são todos de caráter civil e concentrar-se na regulamentação, educação e cooperação.

• Finlândia (2008): A base da estratégia é uma visão de segurança cibernética como uma questão de segurança de dados e como uma questão de importância econômica que está intimamente relacionado com o desenvolvimento da sociedade da informação finlandês.

• Eslováquia (2008): Garantir a segurança da informação é vista como sendo essencial para o funcionamento e desenvolvimento da sociedade. Portanto, o objetivo da estratégia é desenvolver um quadro abrangente. Os objetivos estratégicos estão focados principalmente na prevenção, bem como a disponibilidade e sustentabilidade.

• República Tcheca (2011): os objetivos essenciais da estratégia de segurança cibernética incluem a proteção contra as ameaças que os sistemas e tecnologias de informação e comunicação estão expostos, e mitigação de potenciais consequências em caso de um ataque contra TIC. A estratégia centra-se principalmente em livre acesso aos serviços, a integridade e a confidencialidade dos dados do ciberespaço da República Checa e é coordenado com outras estratégias e conceitos relacionados.

• França (2011): França concentra-se na capacitação de sistemas de informação para resistir a eventos no ciberespaço que poderiam comprometer a disponibilidade, integridade ou confidencialidade dos dados. França salienta os dois meios técnicos relacionados com a segurança dos sistemas de informação ea luta contra a cibercrime eo estabelecimento de um ciberdefesa.

• Alemanha (2011): Alemanha se concentra na prevenção e repressão de ciberataques e também sobre a prevenção de falhas de TI concorrentes, especialmente em infraestruturas críticas. A estratégia define o terreno para a proteção de estruturas de informação críticas. Ele explora os regulamentos existentes para esclarecer se, em caso afirmativo, onde poderes adicionais são necessários para garantir os sistemas de TI na Alemanha por meio de prestação de funções básicas de segurança certificadas pelo Estado e também de apoio às PME através da criação de uma nova força-tarefa.

• Lituânia (2011): a Lituânia tem como objetivo determinar os objetivos e tarefas para o desenvolvimento da informação eletrônica, a fim de garantir a confidencialidade, integridade e acessibilidade da informação eletrônica e serviços prestados no ciberespaço; salvaguarda de redes de comunicações eletrônicas, sistemas de informação e infra-estruturas críticas de informação contra incidentes e ciberataques; protecção dos dados pessoais e privacidade. A estratégia define igualmente as tarefas, que, quando implementadas permitiria a total segurança do ciberespaço e entidades que operam nele.

• Luxemburgo (2011): Reconhecendo a penetração das TIC, a estratégia afirma que é uma prioridade evitar quaisquer efeitos adversos sobre a saúde e a segurança pública ou na economia. Ele também menciona a importância das TIC para os cidadãos, a sociedade e para o crescimento económico. A estratégia baseia-se em cinco linhas de acção. Estes podem resumir-se do PICI e resposta a incidentes; modernização do quadro legal; nacional e da cooperação internacional; educação e conscientização; e promoção de normas.

• Países Baixos (2011): A Holanda tem como objetivo para seguros e confiáveis TIC e medos abuso e (em grande escala) perturbação - e, ao mesmo tempo, reconhece a necessidade de proteger a abertura e a liberdade da Internet. A Holanda incluir uma definição de segurança cibernética na estratégia: "A cibersegurança é estar livre de perigo ou dano causado pela ruptura ou queda-out das TIC ou abuso de TIC O perigo ou o dano devido a abuso, interrupção ou queda-. fora pode ser composto de uma limitação da disponibilidade e confiabilidade das TIC, a violação do sigilo das informações armazenadas no TIC ou danos à integridade das informações. "

• Reino Unido (2011): A abordagem do Reino Unido está a concentrar-se nos objetivos nacionais ligados à evolução segurança cibernética: tornar o Reino Unido a maior economia da inovação, investimento e qualidade na área das TIC e por isso é capaz de explorar todo o potencial e benefícios do ciberespaço. O objetivo é combater os riscos de ciberespaço como os ciber-ataques de criminosos, terroristas e Estados, a fim de torná-lo um espaço seguro para os cidadãos e empresas.

Outras iniciativas

Abaixo está uma breve introdução a três estratégias de países não pertencentes à UE. Muitos outros países também publicaram suas estratégias de segurança, por exemplo, Índia, Austrália, Nova Zelândia e Colômbia. A lista não é exaustiva. No entanto, ela demonstra que a importância da segurança cibernética é reconhecida mundialmente.

• Estados Unidos da América: Os Estados Unidos lançaram a Estratégia Internacional para a Ciberespaço em maio 2011, que descreve um conjunto de atividades em sete áreas interdependentes, baseado em um modelo colaborativo envolvendo o governo, os parceiros internacionais e do setor privado.

• Canadá: a estratégia de segurança cibernética do Canadá foi publicada em 2010 e é construída sobre três pilares: proteger os sistemas de governo; parcerias para proteger os sistemas cibernéticos vitais fora do Governo federal e ajudar os canadenses a ter segurança online. O primeiro pilar visa estabelecer funções e responsabilidades claras, para reforçar a segurança dos sistemas cibernéticos federais e para aumentar a consciência de segurança cibernética em todo o governo. O segundo pilar abrange uma série de iniciativas em parceria com as províncias e territórios e envolver o setor privado e os setores críticos de infraestrutura. Finalmente, o terceiro pilar visa combater o cibercrime e proteger os cidadãos canadenses em ambientes online. As questões de privacidade são notavelmente abordadas neste terceiro pilar.

• Japão: a estratégia de segurança cibernética do Japão de Maio 2010 também pode ser decomposta em uma série de áreas-chave de intervenção: reforço das políticas, tendo em conta possíveis surtos de ciberataques e estabelecimento de uma organização de resposta. Estabelecimento de políticas adaptadas às mudanças no ambiente de segurança da informação. Estabelecer medidas de segurança da informação ativos ao invés de passivos. Os principais pontos de ação abrangidos pela estratégia incluem: superar os riscos de TI para realizar a segurança na vida da nação. Implementação de uma política que fortaleça experiência em segurança nacional e gestão de crises do ciberespaço, e integridade com a política de TIC, como a fundação de atividades socioeconômicas. Estabelecimento de uma política abrangente triádica que abranja os pontos de vista da segurança nacional, gestão de crises, e nação/proteção do usuário. Uma política de segurança da informação com foco no ponto de vista dos usuários do país é particularmente importante. Estabelecimento de uma política de segurança da informação que contribua para a estratégia de crescimento económico. Construir alianças internacionais.

6 CONSIDERAÇÕES

Muito tem sido feito em esforços pela comunidade de Segurança brasileira nos tópicos norteadores. No entanto, esses esforços ainda são muito dispersos e descoordenados. Porém o Brasil possui material humano que não deixa dúvidas de que uma Política de Segurança Cibernética Nacional é de grande alento para a Nação. Desta forma, esta proposta visa consolidar efetivamente os três pilares, pesquisa científica, inovação tecnológica e formação de recursos humanos na área de Segurança Computacional, e contribuir com as TICs no escopo dos Grandes Desafios da Computação no Brasil, como área estratégica indicada no programa TI Maior.

DEDUPLICAÇÃO DE ENTIDADES EM LARGA ESCALA EM PARALELO NO DOMÍNIO DE SAÚDE

Carlos Eduardo S. Pires, Dimas C. do Nascimento Filho, Demetrio Gomes Mestre

Abstract. Entity resolution consists on identifying entities that refer (e.g. patient records) to the same real world object. In health, for instance, this task is useful to find patients registered redundantly on multiple databases. The task consists in comparing each entity from a data set with the other entities of the data set. Since the problem presents a quadratic asymptotic complexity, optimization strategies are commonly used. However, considering the increasing amount of available data, especially in health, even if optimized strategies are used, the problem still presents high a computational cost. This proposal aims to investigate the task of entity resolution in large health databases using a distributed infrastructure. The MapReduce programming model is used to execute optimized entity resolution strategies in parallel. By combining optimization with parallelism we intend to propose still more efficient solutions to the entity resolution problem in the context of large data sets.

Resumo. A deduplicação de entidades consiste em identificar entidades (por exemplo, registros de pacientes) que se referem a um mesmo objeto do mundo real. Na área de saúde, por exemplo, esta tarefa é útil para encontrar pacientes cadastrados de forma redundante em múltiplas bases de dados. A tarefa consiste em comparar cada entidade de um conjunto de dados com as demais entidades do conjunto. Como o problema apresenta uma complexidade quadrática, estratégias de otimização são comumente utilizadas. No entanto, considerando o aumento no volume de dados disponíveis, especialmente no domínio de saúde, mesmo com a aplicação de estratégias de otimização, o problema ainda apresenta um alto custo computacional. Esta proposta visa investigar a deduplicação de entidades em grandes bases de dados de saúde usando uma infraestrutura computacional distribuída. O modelo de programação MapReduce é aplicado para execução em paralelo das estratégias otimizadas. Ao combinar otimização com processamento em paralelo pretende-se propor soluções ainda mais eficientes para o problema da deduplicação de entidades envolvendo grandes conjuntos de dados.

Departamento de Sistemas e Computação
Universidade Federal de Campina Grande (UFCG) – Campina Grande, PB – Brasil

cesp@dsc.ufcg.edu.br, {dimascnf,demetriomestre}@copin.ufcg.edu.br

1. INTRODUÇÃO

Os dados possuem uma importância crescente na atual sociedade da informação e comunicação. Eles são considerados um recurso essencial que possibilita o aumento da produtividade, eficiência e competitividade das organizações, sejam elas públicas ou privadas, independente do nível organizacional (operacional, tático, ou estratégico). No entanto, estima-se que, em média, as organizações possuem entre 1% e 5% de "dados sujos" e que, em algumas delas, essa taxa pode chegar a 30% [Oliveira, 2009]. Dados sujos são dados que apresentam problemas de qualidade devido à existência de erros, inconsistências, redundâncias, entre outros [Sadiq, 2013]. Essas taxas elevadas são responsáveis por prejuízos financeiros e redução da produtividade das organizações, devido ao tempo gasto na adequação dos dados até que possam ser de fato utilizados.

Dentre os principais fatores que contribuem para a baixa qualidade de dados nas organizações, pode-se citar a introdução de dados duplicados, ocasionada normalmente por erro humano. Como um exemplo real, considere uma notícia publicada há dez anos, que na época alertava para as consequências da duplicidade de cadastros: "*Número de registros de cadastros sociais é maior que a população brasileira - Estudo descobre 541 milhões de inscritos, 370 milhões a mais do que o total de habitantes*" [Athias, 2004]. As consequências da duplicação de dados manifestam-se, quotidianamente, de diferentes formas: diminui a usabilidade dos dados, causa gastos desnecessários, gera insatisfação dos clientes, produz indicadores incorretos de desempenho, inibe a compreensão dos dados e de seus valores, entre outras [Ganti & Sarma, 2013]. Para minimizar tais problemas, as organizações necessitam de mecanismos de deduplicação para melhorar a qualidade de seus dados.

A tarefa de deduplicação de entidades (também conhecida na literatura como deduplicação de dados, resolução de entidades, *record linkage*, *entity matching*, *data matching* ou reconciliação de referências) consiste em identificar entidades que se correspondem a um mesmo objeto do mundo real [Christen, 2012]. Um exemplo de entidade pode ser um registro de banco de dados que se refere a um paciente. As entidades podem estar disponíveis em um mesmo conjunto de dados (base de dados) ou em conjuntos de dados distribuídos (em geral, bases de dados heterogêneas). Na comparação de entidades, medidas de similaridade são utilizadas normalmente em conjunto [Koudas et al., 2006], e.g. Jaccard e Cosseno. Tais medidas são aplicadas nos atributos das entidades. Tradicionalmente, a similaridade ou diferença entre duas entidades é determinada por um valor global obtido a partir da combinação dos valores gerados por cada medida individual [Köpcke & Rahm, 2010].

A estratégia tradicional de deduplicação de dados consiste em comparar cada entidade de um conjunto de dados com as demais entidades do mesmo ou de outro conjunto, o que faz com que esta estratégia apresente complexidade quadrática ($O(n^2)$) [Naumann

& Herschel, 2010]. Por exemplo, para um conjunto de dados contendo 1000 entidades, são necessárias 499.500 comparações para determinar quais dessas entidades são duplicadas.

Conforme esperado, uma avaliação de desempenho recente comprovou que a estratégia tradicional de deduplicação de dados apresenta sérios problemas de desempenho [Köpcke et al., 2010]. Para comparar subconjuntos de entidades das bases de dados DBLP[1] e Google Scholar[2] (contendo 2.600 e 64.000 entidades, respectivamente), aplicando o produto cartesiano, foram necessárias 75 horas, considerando um único atributo de cada entidade. Observou-se ainda que o tempo de execução tende a crescer à medida que novos atributos são considerados nas comparações. Nesse sentido, estratégias otimizadas de deduplicação de dados (por exemplo, blocagem padrão, janela fixa, janela adaptativa, janela fixa/adaptativa com múltiplas passagens e incremental) são utilizadas para diminuir o espaço de busca e, consequentemente, alcançar tempos de execução menores.

Entretanto, considerando o aumento no volume de dados disponíveis nas organizações (em particular, nas organizações do setor de saúde), mesmo com a aplicação de estratégias otimizadas, a deduplicação de dados ainda é considerada um problema de alto custo computacional que pode consumir horas ou até mesmo dias. Este problema torna-se ainda mais desafiador caso o resultado da deduplicação necessite ser obtido de forma imediata [Dey et al., 2011]. Diante do exposto, esta proposta visa investigar a tarefa de deduplicação de dados em grandes bases de dados de saúde usando uma infraestrutura computacional distribuída. O paradigma de programação *MapReduce* será utilizado para execução em paralelo de estratégias otimizadas de deduplicação. Ao combinar otimização com processamento em paralelo pretende-se propor uma solução ainda mais eficiente para o problema da deduplicação envolvendo grandes conjuntos de dados.

Atualmente, existem nos sistemas da saúde pública mais de 250 milhões de registros sobre pacientes. Uma parte desses registros é duplicada ou contém algum tipo de problema que os tornam incapazes de identificar um paciente. Combater este cenário é uma tarefa essencial que exige estratégias de deduplicação que foquem no quesito desempenho para poder lidar com grandes volumes de dados. A tarefa de deduplicação de dados pode ser aplicada também no combate ao desperdício financeiro, e.g. departamentos de assistência social, onde há a necessidade de identificar indivíduos que se inscrevem em programas assistencialistas múltiplas vezes ou os indivíduos que trabalham e, ao mesmo tempo, recebem benefícios de auxílio-desemprego.

[1]The DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/~ley/db/
[2]Google Scholar, http://scholar.google.com/

A concretização das contribuições pretendidas por este trabalho poderá ser utilizada por organizações do setor de saúde no intuito de realizar tarefas eficazes, eficientes e contínuas de deduplicação de grandes bases de dados, facilitando assim, a execução de atividades cruciais para as organizações, tais como a detecção de fraudes, integração de dados e tomada de decisões com base em dados confiáveis.

Este documento está organizado da seguinte forma: a seção 2 oferece a fundamentação teórica para uma melhor compreensão da proposta. A seção 3 delimita o escopo da solução proposta. A seção 4 apresenta as contribuições pretendidas. Finalmente, a seção 5 expõe o plano preliminar para desenvolvimento da solução.

2. REFERENCIAL TEÓRICO

Esta seção apresenta a fundamentação teórica necessária para facilitar o entendimento da solução proposta. Primeiramente, são apresentadas as principais estratégias otimizadas para deduplicação de dados. Em seguida, o paradigma de programação MapReduce é introduzido. Por fim, é descrito como pode ser feita a deduplicação de dados seguindo o paradigma MapReduce, além dos principais desafios que podem ser encontrados.

2.1 Estratégias Otimizadas de Deduplicação de Dados

O problema de deduplicação de dados pode ser definido formalmente conforme segue: dados dois conjuntos de entidades $A \in R$ e $B \in S$ das bases de dados R e S, o objetivo é identificar todas as entidades que correspondem ao mesmo objeto do mundo real em A×B. A definição inclui o caso especial de encontrar entidades duplicadas dentro de um única base de dados (A=B; R=S). O resultado da detecção de similaridades é representado por um conjunto de correspondências, onde uma correspondência $c = (e_i; e_j; s)$ relaciona duas entidades e_i e e_j, das bases de dados R e S, segundo um valor de similaridade $s \in [0;1]$ que indica o grau de similaridade entre as duas entidades.

A definição formal enfatiza em sua descrição um dos grandes problemas da deduplicação de dados: "*identificar todas as entidades que correspondem ao mesmo objeto do mundo real em A×B*". O problema está na dificuldade de execução da tarefa de deduplicação de dados, uma vez que existe a necessidade de aplicar medidas de similaridade sobre o produto cartesiano entre as entidades de entrada dos conjuntos A e B. A complexidade quadrática $O(n^2)$, resultante deste processo, indica que a tarefa é bastante ineficiente quando o número de entidades dos conjuntos A e B alcança a ordem dos milhões [Kopcke et al., 2010].

Para reduzir o custo computacional e até mesmo de recursos humanos, é necessário utilizar estratégias otimizadas de deduplicação de dados. Em linhas gerais, tais estratégias visam diminuir o espaço de busca e, consequentemente, alcançar tempos de execução menores. Reduzir o espaço de busca significa eliminar comparações desnecessárias, ou seja, evitar a comparação entre entidades que possuem baixa probabilidade de serem similares. Essa pesquisa considera as seguintes estratégias otimizadas de deduplicação de dados: blocagem padrão, janela fixa, janela adaptativa e janela fixa/adaptativa com múltiplas passagens e incremental.

Na estratégia otimizada baseada em "blocagem padrão" (em inglês, *standard blocking*) é utilizada uma chave de bloco para particionar as entidades a serem comparadas dentro de grupos (blocos). Assim, as entidades só podem ser comparadas com outras entidades que compartilham a mesma chave de bloco, ou seja, a tarefa de deduplicação de dados é executada restritamente dentro de cada bloco. Tipicamente, a chave de bloco de uma entidade é composta por partes dos valores dos atributos da entidade (por exemplo, as primeiras três letras do nome de um paciente concatenadas com as cinco últimas letras do sobrenome).

A estratégia otimizada baseada em "janela de tamanho fixo" (*sorted neighborhood*) é realizada em três fases distintas. Primeiramente, uma chave de ordenamento é atribuída para cada entidade. A chave de uma entidade não precisa ser única em relação às demais. Em geral, é gerada a partir da combinação (concatenação) dos valores dos diferentes atributos da entidade. Em seguida, as entidades são ordenadas de acordo com a chave; assume-se que as entidades duplicadas possuirão chaves iguais ou similares e, por conseguinte, ficarão localizadas próximas entre si após a ordenação. Finalmente, uma janela móvel de tamanho fixo percorre sequencialmente as entidades ordenadas. Os pares de entidades que aparecem dentro da mesma janela são comparados entre si, dois a dois.

Na prática, a distribuição da chave de ordenamento pode não ser uniforme. Em outras palavras, pode haver regiões, no conjunto de entidades, que geram altas ou baixas taxas de detecção de similaridade. Nesse caso, a utilização de uma janela de tamanho fixo pode não ser adequada. Por exemplo, a escolha de uma janela de tamanho pequeno impossibilitará que todas as entidades com uma mesma chave caibam na janela fazendo com que entidades possivelmente similares deixem de ser comparadas. Por outro lado, uma janela de tamanho grande irá provocar a comparação desnecessária de algumas entidades uma vez que, quando a janela for movida, entrarão na janela entidades com chaves iguais e distintas.

Uma alternativa para resolver o problema em questão é utilizar a estratégia otimizada baseada em "janela de tamanho adaptativo" (*adaptive sorted neighborhood*) [Draisbach et al., 2012]. O tamanho da janela é incrementado ou decrementado quando a janela estiver projetada sobre regiões de dados cuja taxa de detecção é alta ou baixa, respec-

tivamente. Tal estratégia promove um aumento na acurácia da detecção de similaridades e uma diminuição significativa do número de comparações realizadas entre entidades. Draisbach et al. provaram, experimentalmente, que a estratégia baseada em janela de tamanho adaptativo supera a estratégia de janela de tamanho fixo em termos de eficiência, apresentando resultados similares em termos de precisão na detecção de similaridades mas realizando um número significativamente menor de comparações.

A utilização de uma única chave de bloco pode não ser suficiente para encontrar todas as entidades duplicadas em um conjunto de dados. Assim, existe ainda uma variação da estratégia baseada em janela conhecida como estratégia baseada em "janela com múltiplas passagens" (*muti-pass sorted neighborhood*) sobre o conjunto de dados [Kolb et al., 2012a]. Essa variante utiliza várias chaves de bloco, geralmente uma chave específica para cada atributo de uma entidade, no sentido de potencializar a taxa de detecção de similaridades.

A deduplicação de entidades é uma tarefa que precisa ser periodicamente executada pelas organizações. Isso ocorre porque as bases de dados, em especial de saúde, são dinâmicas, ou seja, os dados dessas bases são continuamente atualizados, removidos ou adicionados. Desta forma, reexecutar a deduplicação de entidades considerando a totalidade dos dados é algo custoso em termos de consumo de recursos computacionais, mesmo se estratégias otimizadas forem utilizadas. Idealmente, a deduplicação deve ser uma tarefa incremental que envolve apenas a parcela da base de dados contendo entidades que foram modificadas, removidas ou inseridas entre um instante de tempo e outro posterior [Gruenheid et al., 2014; Whang et al., 2014].

2.2 MapReduce

MapReduce (MR) é um modelo de programação desenvolvido para computação intensiva de dados em infraestruturas distribuídas (*clusters*) com grande número de nós [Dean & Ghemawat, 2008]. A ideia chave baseia-se na divisão, distribuição e no armazenamento de dados em um sistema de arquivos distribuído (*Distributed File System* - DFS). As entidades do modelo são representadas por pares chave-valor e a computação é expressa com duas funções principais:

map: $(\text{chave}_{\text{entrada}}, \text{valor}_{\text{entrada}}) \rightarrow \text{lista}(\text{chave}_{\text{tmp}}, \text{valor}_{\text{tmp}})$

reduce: $(\text{chave}_{\text{tmp}}, \text{lista}(\text{valor}_{\text{tmp}})) \rightarrow \text{lista}(\text{chave}_{\text{saída}}, \text{valor}_{\text{saída}})$

Cada uma destas funções pode ser executada em paralelo em partições disjuntas dos dados de entrada. Para cada par chave-valor, a função *map* é chamada e produz, como saída, um par chave-valor temporário que será usado como entrada para a função *reduce*. Diferentemente da função *map*, a função *reduce* é chamada sempre que uma

chave temporária ocorrer como saída de uma função *map*. Entretanto, dentro de uma instância de função *reduce* apenas os valores correspondentes (lista($valor_{tmp}$)) de uma determinada chave podem ser acessados. Um *cluster* MR consiste em um conjunto de nós que executam um número fixo (especificado) de tarefas de *map* e de *reduce*. O mecanismo de alocação específico do *framework* garante que, depois que uma tarefa é terminada, outra tarefa é automaticamente designada para o mesmo processo.

Embora existam vários *frameworks* que implementem o modelo de programação MapReduce, Hadoop [Apache Hadoop, 2014] é a implementação mais popular deste paradigma na comunidade de código aberto. Por esse motivo, as implementações das estratégias de deduplicação apresentadas nesta pesquisa juntamente com as avaliações serão feitas utilizando-se o Hadoop.

Um exemplo do fluxo de dados de uma computação MR é mostrado na Figura 1. O trabalho de MR ilustrado consiste em contar o número de ocorrências dos termos (palavras) de um dado conjunto de entrada (conjunto de documentos), uma tarefa comumente utilizada em Sistemas de Recuperação da Informação [Rijsbergen, 1979]. No exemplo, pode-se perceber que o conjunto de entrada é particionado em dois, de acordo com o número de tarefas de *map* disponíveis, com o intuito de mostrar como a ideia do paralelismo funciona. Apesar do conjunto de dados de entrada representar uma amostra pequena, normalmente as tarefas de *map* processam conjuntos de dados gigantescos.

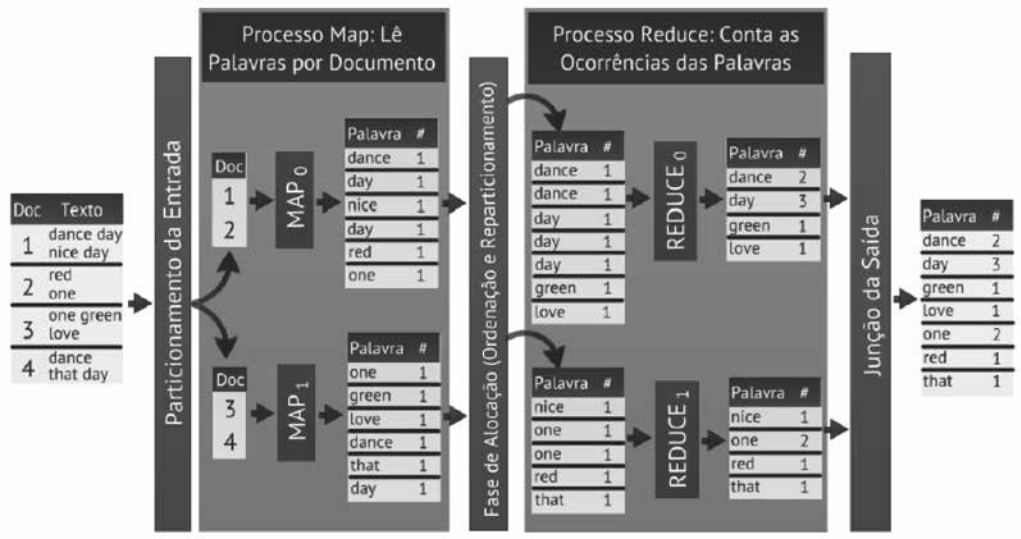


Figura 1. Exemplo de programa MR para contagem da ocorrência de palavras em um conjunto de documentos (adaptado de [Kolb et al., 2012a]).

A Figura 1 mostra que as duas instâncias das funções *map* leem todas as palavras existentes no conjunto de dados e emitem duas listas de pares (termo, 1) particionados por uma função *hash*, conhecida como *part*, com base na chave (Palavra). Depois disso, durante a fase intermediária, entre o processo de *map* e o processo de *reduce*, os pares chave-valor são agrupados por uma função conhecida como *group*, ordenados de acordo com sua chave e enviados através do *cluster* por um procedimento de mistura ou alocação (*shuffle*), o que significa que cada par chave-valor será acessado por uma tarefa de *reduce* apropriadamente alocada. No exemplo, todas as chaves (palavras), começando com letras de *a* a *m*, são enviadas para a primeira tarefa de *reduce* e o restante das chaves são enviadas para a segunda tarefa de *reduce*. Assim, cada instância da função de *reduce* soma e emite o número de ocorrências por chave (palavra).

2.3 Deduplicação de Dados Baseada em MapReduce

As estratégias de deduplicação de dados baseadas em MapReduce (MR), quanto à proposição do modelo MR, podem ser classificadas em simples ou complexas [Kolb et al., 2012a]. Uma estratégia simples executa a tarefa de deduplicação de dados de forma distribuída, sem maiores preocupações com as limitações do modelo MapReduce, ou seja, desconsiderando questões como o balanceamento de carga e os gargalos com gastos de recursos da infraestrutura distribuída. Uma estratégia complexa inclui soluções sofisticadas para mitigar a influência das limitações do modelo MR. Para fins didáticos, neste trabalho será apenas descrita e exemplificada uma estratégia simples, conhecida na literatura como estratégia *Basic* [Kolb et al., 2012b].

Na estratégia *Basic*, a fase *map* determina as chaves de bloco para cada partição e gera como saída uma lista de pares chave-valor (chave de bloco, entidade). Depois, a função de particionamento padrão *hash*, localizada na fase de alocação (*shuffle*), utiliza a chave de bloco para designar os pares chave-valor para as tarefas de *reduce* apropriadas. A fase *reduce* executa a computação das comparações de similaridade geradas a partir do produto cartesiano das entidades por bloco.

A Figura 2 ilustra um exemplo para $n = 9$ entidades de uma base de dados de entrada S utilizando $m = 3$ tarefas de *map* e $r = 2$ tarefas de *reduce*. Note que, primeiramente, na etapa de particionamento, o conjunto de entrada S é dividido em m subconjuntos visando à associação de um subconjunto para cada tarefa de *map* disponível. Depois, cada tarefa de *map* lê os dados em paralelo e determina o valor da chave de bloco K para cada entidade de entrada do seu subconjunto de acordo com a primeira letra da entidade. Assim, as entidades que começam com a letra *A* recebem 1 como valor de chave de bloco, as que começam com a letra *B* recebem 2 como valor de chave de bloco, e assim sucessivamente. Por exemplo, *AND* tem 1 como valor de chave de bloco, pois a entidade começa com a letra *A*.

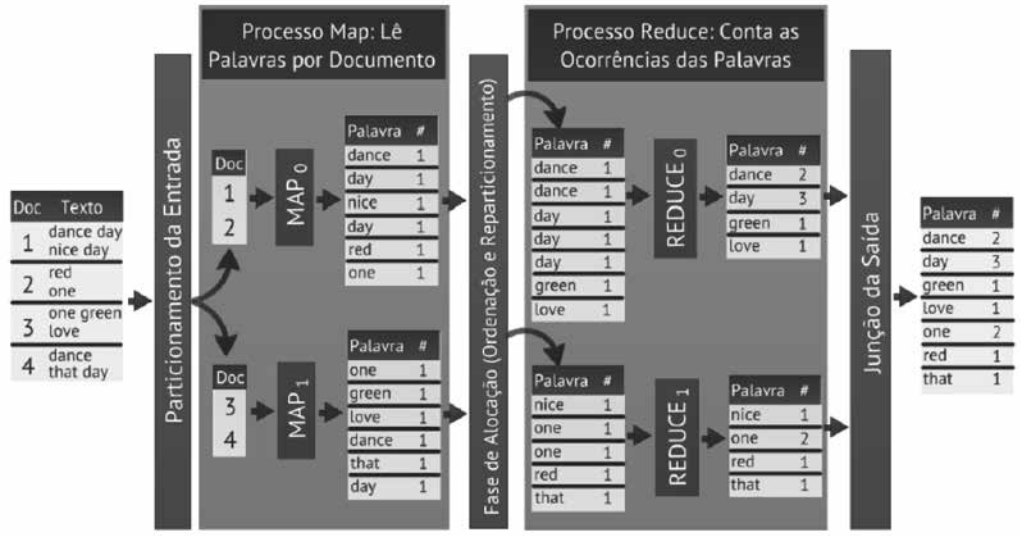


Figura 2. Exemplo de um fluxo de dados de um programa MR para deduplicação de dados (n = 9 entidades de entrada, m = 3 tarefas de *map* e r = 2 tarefas de *reduce*), usando a estratégia *Basic* [Kolb et al., 2012b].

Em seguida, as entidades são dinamicamente distribuídas por uma função de particionamento de tal forma que todas as entidades compartilhando o mesmo valor de chave de bloco sejam enviadas para a mesma tarefa de *reduce*. As entidades com chaves de bloco 1 e 3 são enviadas para a tarefa de *reduce* 0 e as demais entidades são enviadas para a tarefa de *reduce* 1. As funções de *reduce* (que processam tarefas de *reduce*) agrupam as entidades que estão chegando localmente e identificam as duplicações em paralelo. Por exemplo, a tarefa de *reduce* 0 detectou os pares duplicados, marcados com um "*", (AND;AND) e (CIA;CIA). Por fim, as saídas das tarefas de *reduce* são integradas e retornadas como resultado geral da deduplicação de S.

2.3.1. Limitações das Estratégias de Deduplicação de Dados Baseadas em MapReduce

Apesar da implementação de algumas estratégias de deduplicação de dados baseadas em MR parecerem intuitivas, é importante ressaltar que existem pelo menos três limitações inerentes ao modelo MR capazes de deteriorar ou, até mesmo, inviabilizar a execução destas estratégias. Lidar com estas limitações pode aumentar a complexidade das estratégias de deduplicação baseadas em MR que visam obter plena efetividade em termos de desempenho e, por isso, devem ser levadas em consideração sempre que novas estratégias forem propostas. A limitações referem-se a partições disjuntas de dados, balanceamento de carga e gargalos de memória [Kolb et al., 2012b].

Partições Disjuntas de Dados

No paradigma MR, cada função *map* é responsável por ler as entidades pertencentes a uma única partição dos dados e emitir entidades também para uma única partição de saída sem a possibilidade de troca de informações com as demais funções *map* (mesmo as que ainda não entraram em processamento). Tal situação pode gerar complicações no mapeamento das entidades quando existe a necessidade de comparações com entidades de outras partições [Kirsten et al., 2010].

Um exemplo desta limitação pode ser verificado na estratégia de blocagem tradicional, quando as comparações de um bloco grande precisam ser divididas em dois grupos de comparações para serem executados por duas tarefas de *reduce* distintas. É necessário garantir que as entidades envolvidas nas comparações, que serão executadas pela tarefa de *reduce* correspondente, sejam endereçadas apropriadamente para evitar perdas de comparações e, assim, alterar o comportamento da estratégia de blocagem [Mestre & Pires, 2014].

Balanceamento de Carga

O balanceamento de carga está relacionado com a uniformidade da distribuição de carga de trabalho entre duas ou mais tarefas da fase de *reduce* a fim de otimizar a utilização de recursos, maximizar o desempenho, minimizar o tempo de resposta e evitar problemas de sobrecarga. Assim, em termos de balanceamento de carga para a deduplicação de dados baseada em MR, o que se busca é que cada tarefa de *reduce* execute aproximadamente o mesmo número de comparações [Mestre & Pires, 2013]. Para isso, a utilização de mecanismos como a replicação de entidades e a fixação do número de entidades por partições de entrada são geralmente indispensáveis para se alcançar uniformidade na distribuição da carga de trabalho.

A quantidade exata de comparações depende de vários fatores pertinentes ao problema sendo tratado (neste caso, deduplicação de dados), tais como o enviesamento dos dados (diferença nos tamanhos dos blocos de entidades), quantidade de tarefas de *reduce* disponíveis, estratégia de deduplicação de dados a ser utilizada, entre outros. Assim, fica evidente que a solução de balanceamento deve ser tratada na estratégia de deduplicação proposta e, nesse caso, os mecanismos do modelo MR para balanceamento de carga não ajudam na solução. Por exemplo, na sistemática da estratégia *Basic*, dependendo de como é definida a chave de bloco, pode haver a geração de blocos com grandes e pequenos números de entidades provocando, assim, o enviesamento de dados. Então, tendo em vista que todas as entidades de um mesmo bloco são enviadas para uma única tarefa de *reduce*, a computação das comparações dos pares de entidades de blocos grandes pode ocupar por muito tempo alguns nós (da infraestrutura), deixando outros ociosos devido à disparidade (desbalanceamento) do tamanho das tarefas.

Gargalos de Memória

Como mencionado anteriormente, todas as entidades de um mesmo bloco são enviadas para uma única tarefa de *reduce*. Dado que uma tarefa de *reduce* só pode processar dados linha por linha, similar a um cursor SQL, isso implica dizer que todas as entidades de um mesmo bloco precisam ser carregadas na memória antes de serem processadas na fase *reduce*. Assim, a memória de um dado nó pode não ser suficiente para comportar o armazenamento do conjunto de entidades de um bloco muito grande e o processo de MR pode apresentar instabilidade. O problema de memória está fortemente relacionado ao problema de balanceamento de carga, ou seja, ganhos na otimização do balanceamento de carga implicam em ganhos na diminuição de consumo de memória por nó.

3. ESCOPO DA SOLUÇÃO PROPOSTA

Nas últimas décadas, foram desenvolvidos pelo Ministério da Saúde do Brasil importantes sistemas de informação de âmbito nacional. As bases de dados desses sistemas contém uma diversidade de informações – sobre nascimentos, óbitos, doenças de notificação, atendimentos hospitalares e ambulatoriais, atenção básica e orçamentos públicos em saúde, entre outras – que os tornam essenciais para a elaboração de políticas públicas, o planejamento e a gestão, a avaliação e o controle dos serviços de saúde prestados a população, especialmente se os dados puderem ser integrados.

Entretanto, a heterogeneidade no que se refere ao modo de conceber tais sistemas de informação de saúde aliada à ausência de identificadores unívocos para os dados torna a tarefa de integrar as bases de dados bastante complexa. Dado o contexto de criação e desenvolvimento dos sistemas de informação em saúde no Brasil - onde a produção e a utilização de informações se processam em um contexto complexo de relações institucionais, compreendendo mecanismos variados de gestão e financiamento - a integração das diversas bases de dados também ocorre de forma descontínua e desarticulada entre os vários entes da estrutura governamental [Nita et al., 2010].

Na integração de bases de dados, uma das tarefas que necessita ser aplicada é a deduplicação de dados, uma vez que, por exemplo, um mesmo paciente[3] pode estar cadastrado em mais de uma base de dados. No Brasil, a dificuldade em vincular registros pertencentes a um mesmo paciente nas bases de dados disponíveis vem sendo dirimida ao longo dos anos. No entanto, quando se trata de dados de saúde, os números chegam facilmente na ordem dos milhões, como informa o Departamento de Informática do SUS [Datasus, 2014].

[3]Considera-se paciente por ser o elo aglutinador, ou seja, por propiciar a integração entre dados clínicos, assistenciais e administrativos

O problema da duplicação de dados tem sido amplamente estudado por diferentes comunidades de pesquisa, destacando as comunidades de Inteligência Artificial e Banco de Dados, além da própria Indústria [Elmagarmid et al., 2010]. Várias soluções englobando estratégias, *frameworks* e ferramentas (comerciais e gratuitas) foram propostas especialmente no que se refere a dados estruturados [Christen, 2008]. Entretanto, percebe-se que a maioria das soluções existentes não foi concebida para lidar com grandes volumes de dados. A proposta apresentada neste artigo vem no sentido de preencher esta lacuna ao investigar a temática da paralelização da deduplicação de dados em larga escala. Na investigação será considerada a possibilidade dos dados estarem disponíveis em uma única base de dados ou distribuídos em várias bases.

Embora o esforço maior esteja em minimizar o tempo de execução para identificar dados duplicados, a proposta deverá focar também na qualidade do resultado. Neste sentido, deseja-se minimizar a existência de resultados falso-positivos, ou seja, entidades apontadas com duplicadas, mas que de fato são distintas. Quando executada, a solução proposta (em paralelo) deverá realizar as mesmas comparações entre entidades que seriam efetuadas caso a estratégia sequencial de deduplicação correspondente fosse aplicada. Para isso, a solução em paralelo deve garantir que não haja perda de comparações entre entidades o que, nesse caso, afetaria a qualidade do resultado.

Como o conceito de qualidade de dados é transversal a todos os formatos de armazenamento de dados existentes (estruturado, semi-estruturado e não-estruturado), a solução proposta deverá ser genérica, uma vez que as organizações podem armazenar seus dados em qualquer formato.

4. CONTRIBUIÇÕES PRETENDIDAS

A solução proposta para melhorar o desempenho da deduplicação de entidades envolvendo grandes bases de dados prevê as seguintes contribuições: proposição de estratégias otimizadas de deduplicação que explorem o paralelismo e considerem o caráter dinâmico das bases de dados, além de um mecanismo para a provisão dinâmica de recursos computacionais em nuvem sobre os quais as estratégias de deduplicação devem ser executadas. Tais contribuições são detalhadas a seguir.

4.1. Estratégias Otimizadas de Deduplicação de Dados utilizando MapReduce

A seguir, são detalhadas as estratégias otimizadas de deduplicação de dados a serem proposta.

Estratégia baseada em Blocagem

Visa aprimorar o estado da arte do método de blocagem padrão baseado em MapReduce (MR) com balanceamento de carga. Para tanto, esta contribuição consistirá em um pro-

cessamento composto por dois trabalhos de MR. Em linhas gerais, o primeiro trabalho de MR coleta e armazena, em uma Matriz de Distribuição de Blocos (*Block Distribution Matrix* - BDM), informações acerca da distribuição das entidades, blocos e partições de entrada. O segundo trabalho realiza o balanceamento de carga na fase *map* e as comparações entre as entidades na fase *reduce* com base nas estatísticas armazenadas na BDM.

Estratégia baseada em Janela Fixa com Passagens Simples e Múltiplas

Visa aprimorar o estado da arte da estratégia *Sorted Neighborhood* (SN) baseada em MR com balanceamento de carga. Assim como na estratégia de blocagem padrão, esta contribuição consistirá em um processamento composto por dois trabalhos de MR. O primeiro trabalho de MR coleta e armazena, em uma Matriz de Partições de Chaves (*Key Partitioning Matrix* - KPM), informações acerca da distribuição das entidades por chave de bloco pertencentes a cada uma das partições de entrada. No segundo trabalho de MR, o balanceamento de carga é realizado na fase *map* com base nas estatísticas disponíveis na KPM, tanto para o caso de execução do SN com passagem simples (i.e., elaboração da chave de bloco em relação a apenas um dos atributos da entidade) quanto para o caso da execução do SN com múltiplas passagens (e.g., elaboração de uma chave de bloco para cada atributo da entidade). A descida da(s) janela(s) para a execução das comparações entre as entidades é realizada na fase *reduce*.

Estratégia baseada em Janela Adaptativa com Passagens Simples e Múltiplas

A proposição desta estratégia baseada em MR é inédita na literatura. Trata-se da paralelização de uma estratégia baseada em janelas que se adaptam de acordo com a taxa de detecção de similaridade. É importante ressaltar que é comum, quando lidamos com deduplicação baseada em MR, definir uma estratégia eficiente (com mecanismos de balanceamento de carga) quando sabemos previamente quais serão as comparações a serem realizadas com a descida de uma janela de tamanho fixo. No entanto, a relevância em responder a pergunta de pesquisa, i.e., se é possível a implementação de uma estratégia baseada em MR da janela adaptativa com balanceamento de carga, está na imprevisibilidade da definição das comparações que serão executadas, tendo em vista que o tamanho da janela (ou seja, o número de comparações) se adapta de acordo com a taxa de detecção de entidades duplicadas. A pergunta é: como designar comparações de entidades para as tarefas de *reduce* de forma eficiente sem saber previamente todas as comparações que devem ser realizadas?

Para tentar responder a essa pergunta de pesquisa, assim como nas estratégias anteriores, a contribuição para o SN com janela adaptativa baseado em MR será propor uma tarefa de deduplicação composta por dois trabalhos de MR. O primeiro trabalho de MR analisa e define quais partições de entrada devem ser agrupadas na fase de descida da janela de forma que a janela adaptativa possa crescer e realizar as comparações sem perdas (em relação à execução da estratégia sem mecanismos de paralelização). Os

agrupamentos são armazenados na Matriz de Alocação de Partições (*Partition Allocation Matrix* – PAM). A estrutura da PAM é planejada de forma a comportar também as informações necessárias para execução de múltiplas passagens da janela adaptativa. No segundo trabalho de MR, as partições são designadas para as tarefas de *reduce* apropriadas de acordo com as informações contidas na PAM. A descida da(s) janela(s) adaptativa(s) juntamente com as comparações é realizada na fase *reduce*.

Estratégia Incremental

O objetivo de uma estratégia incremental de deduplicação de entidades é avaliar (i.e., realizar a comparação entre entidades) apenas um subconjunto da base de dados D após subsequentes conjuntos de alterações (ΔD_1, ΔD_2, ... , ΔD_n) sobre D. Primeiramente, é realizado um agrupamento (*clustering*) das entidades de cada bloco (assumindo que uma estratégia de blocagem já teria sido aplicada) e, a partir do conjunto de grupos (*clusters*) resultantes do agrupamento, as entidades que pertençam a um mesmo *cluster* são classificadas como duplicadas. É necessário armazenar metadados sobre a base de dados, as entidades duplicadas, além do grau de acoplamento entre *clusters* e a estrutura atual dos *clusters*.

Após a realização de futuras alterações (ΔD_i) sobre a base de dados, apenas a porção da base de dados D que foi de fato influenciada por ΔD_i é reprocessada para a atualização da estrutura e acoplamento entre os clusters e, por consequência, a identificação das entidades duplicadas.

Uma estratégia incremental de deduplicação de entidades deve apresentar três características principais: a) obter resultados iguais ou similares, em relação ao número de entidades duplicadas detectadas e a porcentagem de falso positivos reportados, em relação à estratégia não incremental; b) executar significativamente mais rápido do que a estratégia não incremental; e c) utilizar informações relativas às atualizações na base de dados no intuito de identificar e corrigir erros previamente cometidos devido à falta de acurácia nos valores de dados anteriormente processados.

Desse modo, são contribuições esperadas deste trabalho: i) a proposição e avaliação de heurísticas para a determinação do subconjunto de uma base de dados que deve ser reprocessado (i.e., que deve ser avaliado para a detecção de entidades duplicadas) após ser realizado um conjunto de atualizações sobre esta base; e ii) a proposição e avaliação de uma estratégia incremental de deduplicação de dados em paralelo.

4.2 Mecanismo de Provisionamento de Recursos para Deduplicação de Dados na Nuvem

Além da execução em paralelo de estratégias otimizadas de deduplicação de dados utilizando, por exemplo, uma infraestrutura baseada em Hadoop, outra maneira possível

de processar grandes volumes de dados em tempo hábil é por intermédio da provisão dinâmica de recursos computacionais [Badidi, 2013]. Tal solução é usualmente realizada com a utilização de tecnologias de computação em nuvem [Schnjakin et al., 2010].

A computação em nuvem emergiu recentemente como um modelo de computação para permitir, de maneira conveniente, ubíqua e sob demanda, acesso via rede a uma séria de recursos computacionais (serviços de rede, máquinas virtuais, armazenamento e aplicações) configuráveis que podem ser rapidamente provisionados e disponibilizados com o mínimo de esforço de gerenciamento e interação com o provedor do serviço de computação em nuvem. Dessa maneira, os recursos computacionais em uma nuvem podem ser aumentados no intuito de tratar processamentos intensos em janelas de tempo específicas ou requisições de múltiplos usuários a um mesmo recurso compartilhado. De maneira análoga, os recursos podem ser desalocados uma vez que demandas específicas sejam atendidas.

O paradigma de computação em nuvem pode ser utilizado para prover uma infraestrutura, com capacidade de provisão dinâmica de recursos, no intuito de processar tarefas de deduplicação de entidades que demandem esforço de processamento intensivo. Nesse contexto, uma contribuição deste trabalho é propor um mecanismo capaz de estimar o esforço computacional das tarefas e, com base nas estimativas, alocar dinamicamente uma infraestrutura computacional para a execução das estratégias de deduplicação de maneira distribuída. Dessa forma, além de melhorar o desempenho da tarefa de deduplicação, a utilização de algoritmos de provisionamento eficazes pode também minimizar o custo (de infraestrutura) para o processamento de grandes bases de dados que necessitem cumprir requisitos não funcionais relacionados ao tempo de execução pré-estabelecidos.

5. PLANO PRELIMINAR PARA DESENVOLVIMENTO DA SOLUÇÃO

As estratégias otimizadas de deduplicação de entidades serão especificadas, desenvolvidas e validadas utilizando uma metodologia incremental de desenvolvimento de software. As limitações do modelo MapReduce (partições disjuntas de dados, balanceamento de carga e gargalos de memória) serão tratadas já na fase de concepção das estratégias otimizadas, a serem propostas na seguinte ordem: blocagem, janelas fixa, janela adaptativa, janela com múltiplas passagens e incremental. Um procedimento semelhante de desenvolvimento de software será aplicado na construção do mecanismo de provisionamento dinâmico de recursos para a execução das estratégias de deduplicação de maneira distribuída. Durante o desenvolvimento da solução, todas as especificações serão devidamente documentadas em relatórios técnicos e manuais.

O desenvolvimento será realizado utilizando a linguagem de programação Java e o *framework* Hadoop. Consideraremos inicialmente que as entidades estarão armazenadas

em bases de dados estruturadas, porém temos a pretensão de estender a proposta para lidar com dados representados em outros formatos. Para validação das estratégias otimizadas de deduplicação de entidades e do mecanismo de provisionamento de recursos, consideraremos diferentes cenários, bases de dados reais do domínio de saúde, além das medidas de similaridade mais utilizadas nos trabalhos relacionados. Parcerias com organizações governamentais deverão ser estabelecidas no sentido de ter acesso às bases de dados reais.

REFERÊNCIAS BIBLIOGRÁFICAS

Apache Hadoop (2014). http://hadoop.apache.org/, último acesso em 08/Setembro/2014.

Athias, G. (2004). http://www1.folha.uol.com.br/fsp/brasil/fc2403200427.htm, último acesso em 11/Setembro/2014.

Badidi, E. (2013). A Cloud Service Broker for SLA-based SaaS provisioning. In Proc. of the International Conference on Information Society (i-Society), pp. 61–66.

Christen, P. (2008) Febrl: a freely available record linkage system with a graphical user interface. In Proceedings of the 2nd Australasian Workshop on Health data and knowledge management – Vol. 80, HDKM'08, pp. 17-25.

Christen, P. (2012) Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Data-Centric Systems and Applications series.

Datasus. (2014) http://www.telesintese.com.br/datasus/, último acesso em 08/Setembro/2014.

Dean, J., Ghemawat, S. (2008) MapReduce: simplified data processing on large clusters. Commun. ACM, 51(1):107–113.

Dey, D., Mookerjee, V., Liu, D. (2011) Efficient Techniques for Online Record Linkage. IEEE Transactions on Knowledge and Data Engineering, 23(3):373-387.

Draisbach, U., Naumann, F., Szott, S., Wonneberg, O. (2012) Adaptive windows for duplicate detection. In Proceedings of the IEEE 28th International Conference on Data Engineering, ICDE'12, pages 1073–1083.

Elmagarmid, A. K., Ipeirotis, P. G. Verykios, V. S. (2007) Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19(1):1-16.

Ganti, V., Sarma, A. D. (2013) Data Cleaning: A Practical Perspective, Morgan & Claypool Publishers.

Gruenheid, A., Dong, X. L., Srivastava, D. (2014) Incremental Record Linkage. Proceedings of the VLDB Endowment, 7(9):697-708.

Kirsten, T., Kolb, L., Hartung, M., Gross, A., Kopcke, H., Rahm, E. (2010) Data Partitioning for Parallel Entity Matching. In Proc. of the 8th International Workshop on Quality in Databases.

Kolb, L., Thor, A., Rahm, E. (2012) Load balancing for mapreduce-based entity resolution. In Proceedings of the 28th International Conference on Data Engineering, ICDE'12, pp. 618–629.

Kolb, L., Thor, A., Rahm, E. (2012) Multi-pass sorted neighborhood blocking with mapreduce. Comput. Sci., 27(1):45–63.

Köpcke, H., Rahm, E. (2010) Frameworks for entity matching: A comparison. Data Knowledge Engineering, 69(2):197-210.

Köpcke, H., Thor, A., Rahm, E. (2010) Evaluation of entity resolution approaches on real-world match problems. Proc. VLDB Endow., 3(1-2):484–493.

Koudas, N., Sarawagi, A., Srivastava, D. (2006) Record Linkage: Similarity Measures and Algorithms. In Proc. of the ACM SIGMOD International Conference on Management of Data, pp. 802-80.

Mestre, D. G., Pires, C. E. S. (2013) Improving Load Balancing for MapReduce-based Entity Matching. In 18th IEEE Symposium on Computers and Communications (ISCC'13), Split, Croatia.

Mestre, D. G., Pires, C. E. S. (2014) Efficient Entity Matching over Multiple Data Sources with MapReduce. Journal of Information and Data Management, 5(1):40-51.

Naumann, F., Herschel, M. (2010) An Introduction to Duplicate Detection. Morgan and Claypool Publishers.

Nita, M. E., Secoli, S. R., Nobre, M. R. C, Ono-Nita, S. K., Campino, A. C., Santi, F. M., Costa, A. M. N., Carrilho, F. J. (2010) Avaliação de Tecnologias em Saúde: Evidência Clínica, Análise Econômica e Análise de Decisão. Ed. Artmed, 1ª Edição.

Oliveira, P. J. M. (2009) Detecção e Correção de Problemas de Qualidade dos Dados: Modelo, Sintaxe e Semântica. Tese de Doutorado. Universidade do Minho, Portugal.

Rijsbergen, C. J. (1979) Information Retrieval, 2nd Edition Stoneham, MA: Butterworths, http://www.dcs.gla.ac.uk/Keith/Preface.html.

Sadiq, S. (2013) Handbook of Data Quality. Research and Practice. Springer, 2013.

Schnjakin, M., Alnemr, R., Meinel, C. (2010) Contract-based Cloud Architecture. In Proc. of the 2nd International Workshop on Cloud Data Management, pp. 33–40.

Whang S. E., Garcia-Molina, H. (2014) Incremental Entity Resolution on Rules and Data. VLDB Journal, 23(1):77–102.

BIOMETRIA POR PADRÕES PAPILOSCÓPICOS, EXPRESSÕES FACIAIS E GESTOS COM APLICAÇÃO EM SAÚDE E SEGURANÇA

Olga R. P. Bellon e Luciano Silva

Abstract. We present the research challenges of IMAGO Group on biometrics. We have been working on this subject since 2005, resulting in publications, awards, patent, theses and dissertations. Our focus is on image acquisition, processing and analysis and the main goals are: 1) facial expressions, gesture recognition to identify emotions and behavior, and 2) neonatal and adult fingerprint recognition, seeking for fraud robustness, to be used at hospitals, borders, and disasters. The combination of the IMAGO's 20 years expertise on depth (3D) images and the new image acquisition technologies (2D, 3D) has produced innovations on biometrics applied to Health Care and Security.

Resumo. Apresentamos os desafios de pesquisas em biometria do Grupo IMAGO. Temos atuado neste tema desde 2005, produzindo publicações, premiações, patente, teses e dissertações. Nosso foco está na aquisição, processamento e análise de imagens, e os objetivos principais são: 1) reconhecer expressões faciais e gestos, para identificar sentimentos e comportamentos; e 2) identificar digitais neonatal e de adultos, buscando eliminar a possibilidade de fraude, para uso em hospitais, fronteiras, e casos de desastres. A união da experiência de 20 anos do grupo em imagens de profundidade (3D) e as novas tecnologias de aquisição de imagens (2D e 3D) tem produzido inovações em biometria, com aplicação em Saúde e Segurança.

1. INTRODUÇÃO

Este trabalho apresenta nossas contribuições de pesquisa em biometria, bem como as bases, justificativas e desafios dos trabalhos recém-iniciados. O objetivo é mostrar a relevância das pesquisas e sua aderência aos Grandes Desafios da Computação, com aplicação em Saúde e Segurança e forte apelo Social. O grupo IMAGO-UFPR tem pesquisado padrões papiloscópicos, faces, e, mais recentemente, gestos. Nosso objetivo foi sempre contribuir para a solução de problemas nacionais, bem como dominar e desenvolver tecnologias de ponta no assunto. O primeiro problema abordado foi a identificação neonatal, visando eliminar situações de troca, roubo, ou tráfico. O tema foi escolhido após verificar que os "borrões", obrigatoriamente coletados nas maternidades, não

Grupo IMAGO de Pesquisa – Universidade Federal do Paraná
Caixa Postal 19092, CEP 81531-980, Curitiba – PR – Brasil

{olga,luciano}@ufpr.br www.imago.ufpr.br

possuem qualidade mínima para identificação. Essa pesquisa original passou por diversos momentos [e.g., 1-3]. Foi comprovado por papiloscopista do IITB (PE) que as impressões palmares, coletadas com o sensor desenvolvido grupo, podem ser utilizadas na prática para identificação neonatal. Outro problema abordado foi o reconhecimento biométrico facial. Exploramos nosso método de registro [4] aplicado ao reconhecimento de faces [5]. Uma nova fase foi iniciada recentemente com imagens RGB-D [6].

2. TRABALHOS EM ANDAMENTO

Até o momento, pesquisamos em detalhes impressões digitais de neonatos. O objetivo é aplicar a mesma biometria usada em adultos, para uniformizar a identificação do cidadão. Digitais de neonatos têm características que impõem uma série de dificuldades adicionais, se comparadas com as de adultos [1]. Usamos nosso sensor para coletar 350 imagens de neonatos, e a qualidade das imagens para identificação foi atestada por papiloscopistas. O próximo passo é coletar estas imagens em larga escala, e desenvolver uma metodologia que garanta uma identificação definitiva para os neonatos, com base em nossas técnicas para as superfícies palmar e plantar. Finalmente, já contamos um primeiro resultado preliminar no desenvolvimento de sensor baseado em OCT [7], que permite a obtenção de imagens subcutâneas. Este sensor podendo ser usado inclusive em tecidos deteriorados, situação comum em desastres. Colaboração: Prof Anil Jain (MSU-USA), Prof Audrey Ellerbee (Stanford-USA), Mater. Victor Amaral, IITB (PE).

Sobre reconhecimento de faces, iniciamos um estudo sobre identificação de expressões de dor em neonatos, para melhorar seu atendimento e reduzir as chances de óbito ou sequelas. Algumas iniciativas já foram realizadas para comprovar que os neonatos sentem dor, mas são limitadas em termos de análise de imagens. Nosso trabalho pretende não apenas avaliar e quantificar as expressões, mas também qualificá-las. Ainda sobre faces, e incluindo agora gestos, montamos um cenário para coleta de uma grande base de imagens e vídeos (2D e 3D) no HC-UFPR. A coleta desta base terá que ser corretamente padronizada para permitir pesquisas sobre diversas desordens de movimento, e.g., mal de Parkinson e autismo. Estas desordens ocorrem em diversas parcelas da população, incluindo usuários de crack (Parkinson) e neonatos destas usuárias (autismo); serão identificadas expressões e gestos típicos destes problemas. Colaboração: Dr Hélio Teive (neurologista HC), Dr Gustavo Dória (pediatria HC), prof Dmitry Goldgof e Sudeeep Sarkar (USF-USA), Ruth Guinsburg (pediatria UNIFESP).

3. CONCLUSÃO

Apresentamos pesquisas desafiadoras para a solução de grandes problemas nacionais, com aderência aos Grandes Desafios. Além das contribuições em Computação, há também contribuição Social. As pesquisas são multidisciplinares.

REFERÊNCIAS

[1] D. Weingaertner, O.R.P. Bellon, L. Silva. Newborn´s biometric identification: can it be done?. Intl Conf. on Computer Vision Theory and Applications (VISAPP), 2008.

[2] R.P. Lemes, O.R.P. Bellon, L. Silva. Dynamic pore filtering for keypoint detection applied to newborn authentication. Intl Conf. on Pattern Recognition (ICPR), 2014.

[3] R.P. Lemes, O.R.P. Bellon, L. Silva, A. Jain. Biometric Recognition of Newborns: Identification using Palmprints. Intl Joint Conference on Biometrics, 2011.

[4] L. Silva, O.R.P. Bellon, K. Boyer. Precision image registration using surface interpenetration measure enhanced genetic algorithms. IEEE TPAMI, 27:762, 2005.

[5] C.C. Queirolo, L. Silva, O.R.P. Bellon, et al. 3D face recognition using simulated annealing and the surface interpenetration measure. IEEE TPAMI, 32:206-219, 2010.

[6] M. Pamplona Segundo, L. Silva, O.R.P. Bellon, S. Sarkar. Orthogonal projection images for 3D face detection. Pattern Recognition Letters, 2014.

[7] H.S.G. Costa, L. Silva, A.K. Ellerbee. Evaluation of fingerprint deformation using optical coherence tomography. SPIE Photonics, 2014.

OTIMIZAÇÃO DE COMPONENTES EM SISTEMAS INTEGRADOS VISANDO REDUZIR O CONSUMO DE ENERGIA

Ricardo Reis

Abstract.

This paper is related to the design of low power integrated systems. To reach this goal the power reduction must be considered in all levels of abstractions in the synthesis flow. As the nanoelectronics technologies have a high static power consumption that is related to the amount of transistors it is fundamental to optimize the amount of transistors used to do a chip. The design of chips for medical applications, specially the ones to be implanted in humans, demands the use of energy consumption optimization techniques.

Resumo.

Este artigo trata do projeto de sistemas integrados que tenham baixo consumo de energia. Para tanto a redução do consumo deve ser tratada em todos os níveis de abstração de síntese do sistema integrado. Como as tecnologias de nanoeletrônica atuais possuem um alto consumo estático, que está relacionado com o número de transistores, é fundamental otimizar o número de transistores utilizados na geração de um chip. A realização de chips para aplicações na área médica, especialmente os chips usados em sistemas implantados no ser humano, demanda o uso de técnica de otimização do consumo de energia.

1. INTRODUÇÃO

No contexto dos Grandes Desafios em Computação [SBC 2006], definidos pela SBC, este artigo refere-se ao desafio "Impactos para a área da Computação da transição do silício para novas tecnologias" aplicado ao domínio da saúde. Na realidade este desafio tem relação também com os demais domínios de interesse listados na chamada (sistema bancário, energia, defesa cibernética, educação), mas também com outros domínios não listados, como o aeroespacial e automotivo. Mas neste texto focamos especialmente o domínio da saúde. Na saúde encontramos um número expressivo de aplicações da computação de demandam soluções com baixo consumo e com alta confiabilidade, o que muitas vezes demanda uma solução que use um circuito integrado de aplicação específica (ASIC).

PGMicro/PPGC, Instituto de Informática – Univ. Federal do Rio Grande do Sul
Porto Alegre - Brasil
reis@inf.ufrgs.br

Não apenas na medicina, mas em diversas outras áreas, um dos requisitos cada vez mais importantes, no projeto de sistemas eletrônicos/computacionais integrados, é a redução de consumo. Em sistemas complexos com um número expressivo de componentes, estamos atingindo uma etapa denominada de "Power Wall". Ou seja, limitação em incluir mais componentes (que podem permitir executar novas funcionalidades) devido ao "chip" consumir além do que consegue dissipar de energia. Isto leva à uma pesquisa cada vez maior visando a obtenção de sistemas com alto grau de otimização em todos os níveis de abstração de projeto. Quando tratamos do nível de síntese física, devemos ter novas metodologias que permitam reduzir o número de componentes. Um primeiro passo é colocar todo o sistema computacional/eletrônico em um único chip. O segundo é otimizar o número o transistores necessários para implementar uma função. Isto porque nas nanotecnologias CMOS, o consumo estático passa a superar o consumo estático, e este consumo estático está relacionado ao número de transistores.

Queremos observar que em 2006 a SBC publicou um texto indicando os 5 grandes desafios para a pesquisa avançada em Computação até 2016 [SBC 2006]. Este artigo está relacionado ao desafio "Impactos para a área de computação da transição do silício para novas tecnologias" mas considerando o dito na publicação que fizemos em [Bampi 2009] de que esta transição não deverá ocorrer nas próximas décadas e que a tecnologia CMOS ainda possui ao menos muitas décadas de vida. Esta afirmação também está amparada no ITRS (International Technology Roadmap for Semiconductors) [ITRS 2009], que claramente mostra que a tecnologia CMOS ainda será a tecnologia principal para a implementação de sistemas integrados durante várias décadas. Cabe observar que este roadmap é elaborada por especialistas de empresas, centros de pesquisa e universidades que estão na ponta dos avanços tecnológicos do setor de semicondutores. Estes roadmaps vem sendo desenvolvidos desde 1992, tendo sido demonstrado ao longo do tempo uma significativa precisão, mas os avanços da tecnologia CMOS tem sido mais significativos do que os previstos nos primeiros roadmaps.

2. APLICAÇÕES NA SAÚDE

Quando tratamos de sistemas integrados para aplicação na medicina/saúde, seja sistemas implantados no organismo humano, quando usados externamente, como equipamentos de monitoração da saúde, dentre outros, temos dois grandes requisitos: miniaturização e baixíssimo consumo. Cabe observar que a redução de consumo além de aumentar o tempo para consumir a energia de uma bateria, também contribui para aumentar a confiabilidade do sistema. Sistemas/circuitos integrados podem sofrer falhas devido à existência de regiões de alto consumo/dissipação de energia. Portanto, a redução de consumo contribui também para diminuir o risco de falhas em sistemas integrados que sejam implantados no organismo humano.

Nossa equipe de pesquisa tem se dedicado, há tempos, à pesquisa de técnicas de otimização do número de transistores e automação do leiaute de qualquer rede de transistores. Estes trabalhos tem tido repercussão internacional, como pode ser observado pelo número de artigos publicados no tema, assim como pelo número de palestras convidadas internacionais realizadas nos últimos anos (isto pode ser observado no CV Lattes).

Além de diminuir o número de transistores, é importante também efetuar uma otimização do dimensionamento dos mesmos. O trabalho desenvolvido por nossa equipe, no tema de dimensionamento automático de transistores (gate sizing) venceu o concurso de ferramentas de EDA organizado pela Intel, junto ao International Symposium on Physical Design, ISPD, em 2013. Este trabalho tem continuado, tendo proporcionado colaboração e interesse da IBM Austin e da Synopsys. Os sistemas no domínio da medicina precisam ter um bom dimensionamento de seus componentes de forma que possamos reduzir o consumo de energia do sistema.

Outra pesquisa de importância para sistemas integrados implantáveis em seres humanos, trata do aumento do tempo de vida destes sistemas. As trilhas metálicos em circuitos integrados, podem sofrer efeitos de eletromigração ao longo de um fio metálico, devido à intensidade de correntes, ocasionando "curtos circuitos" ou rompimento de trilhas. Uma de nossas pesquisas (em colaboração com Univ. de Minnesota) trata da pesquisa de técnicas de leiaute para reduzir as densidades de corrente nas linhas metálicas e com isto aumentar o tempo de vida de um circuito/sistema integrado [Posser 2014].

Consideramos as pesquisas descritas sucintamente acima, são estratégicas para que possamos desenvolver, no Brasil, sistemas integrados implantáveis que tenham longa vida (sem demandar nova cirurgia para troca do sistema), que tenham baixo consumo e alta confiabilidade.

Mas um objetivo importante é a obtenção de sistemas otimizados quanto ao consumo e que consigam retirar a energia para o seu funcionamento do ambiente e/ou do movimento do ser humano. Especialmente no projeto de sistemas integrados a serem implantados em seres humanos, é desejado que os mesmos possam usar a energia do ambiente em que estão inseridos (calor humano, movimento humano). Portanto, ao mesmo tempo em que estes sistemas (chips) devem ser projetados usando técnicas de ultra baixo consumo, devem ter circuitos que consigam retirar energia do ambiente onde estão implantados. Uma fonte de energia é a temperatura do ser humano, outra fonte é a geração de energia pelo movimento da pessoa. A área de pesquisa que visa obter energia do ambiente é denominada de "Colheita de Energia" (Energy Harvesting), e deve contribuir para a geração de sistemas implantáveis que possam funcionar sem baterias e portanto evitando que o paciente tenha de tempos em tempos efetuar nova cirurgia para troca das baterias.

3. TRANSISTORES E ENERGIA

O custo de fabricação de cada transistor em sistema de alta escala de integração tem sido cada vez menor. Se compararmos com o custo de grão de arroz, conforme publicado na edição de 6 de setembro de 2010 da revista "The Economist" [Economist 2010], o custo de um grão de arroz pode ser equivalente ao custo de fabricação de mais de 100 mil transistores. Isto poderia indicar que não temos necessidade de economizar o número de transistores em um projeto, já que o custo deles é relativamente pequeno. Porém o custo da energia necessária para a operação de um transistor é cada vez mais elevado. Também temos de considerar que um alto consumo de potencia pode reduzir a vida útil de uma sistema, assim como aumentar os efeitos de variabilidade que podem provocar um mau funcionamento de um sistema integrado.

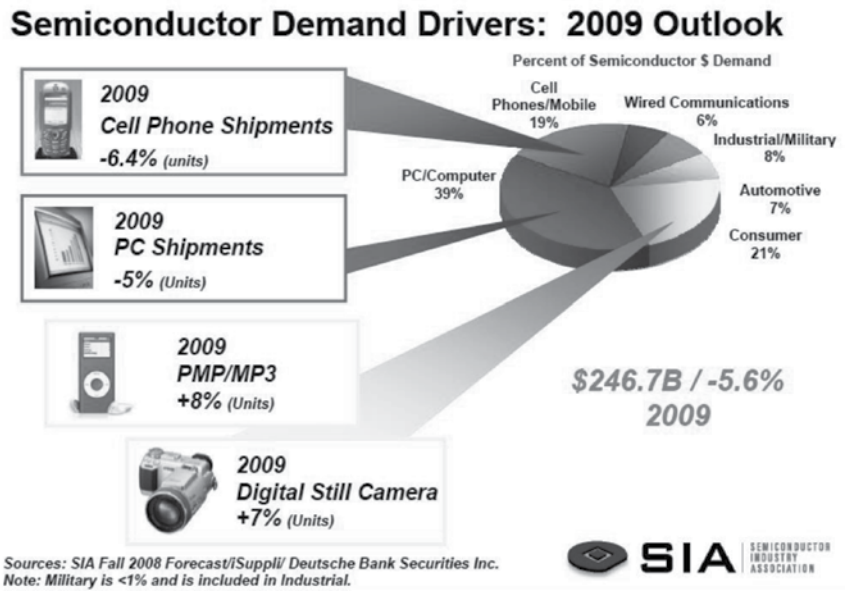


Figura 1 - Demanda de Semicondutores no Mercado Mundial em 2009 (SIA 2009)

O rápido progresso no número de transistores passíveis de serem integrados em um único chip e o fato de que tecnologias CMOs não tinham consumo estático que contribuísse para o consumo global do circuito, provocou a consolidação de metodologias de síntese lógica [Brayton 1984] [Jacobi 1996] e síntese física que não efetuam muita atenção à redução do número de transistores de um circuito. A metodologia baseada no uso de células–padrão (standard cell) obtidas de uma biblioteca de células conduz à uma etapa de síntese lógica que tem de transformar as equações lógicas para equações

que tenham termos correspondentes às funções lógicas das células existentes em uma biblioteca de células [Reis 2009].

Na Figura 2 [Reis 2011] podemos ver o exemplo de uma pequena equação implementada com 4 portas lógicas, com um total de 14 transistores. Na Figura 3 a mesma equação da Figura 2 é implementada usando apenas uma porta lógica, com um total de 8 transistores, ou seja, executada a mesma função mas com um número muito menor de transistores. Além disto reduz em 3 o número de conexões entre portas lógicas a serem efetuadas em um dos níveis de metal. Portanto, reduz também o número de contatos e vias [Figura 4].

Ou seja, o uso de portas lógicas que implementam funções complexas em vez destas funções serem implementadas por um conjunto de portas lógicas básicas, permite reduzir de forma expressiva o número de transistores e consequente o consumo estático.

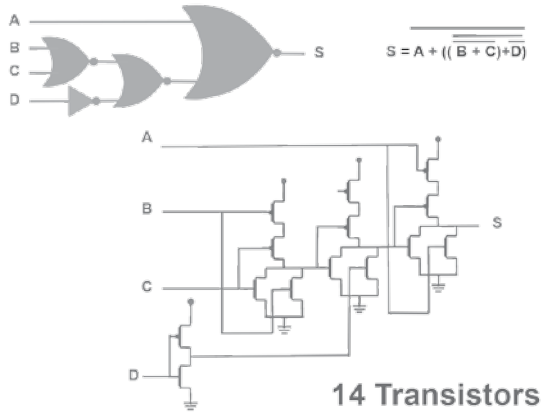


Figura 2 - Circuito implementado com 4 portas lógicas, com um total de 14 transistores

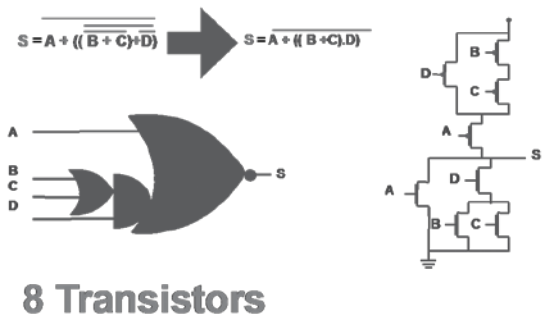


Figura 3 - Circuito implementado com apenas uma porta lógica, com um total de 8 transistores, correspondendo a mesma função do circuito da Figura 2.

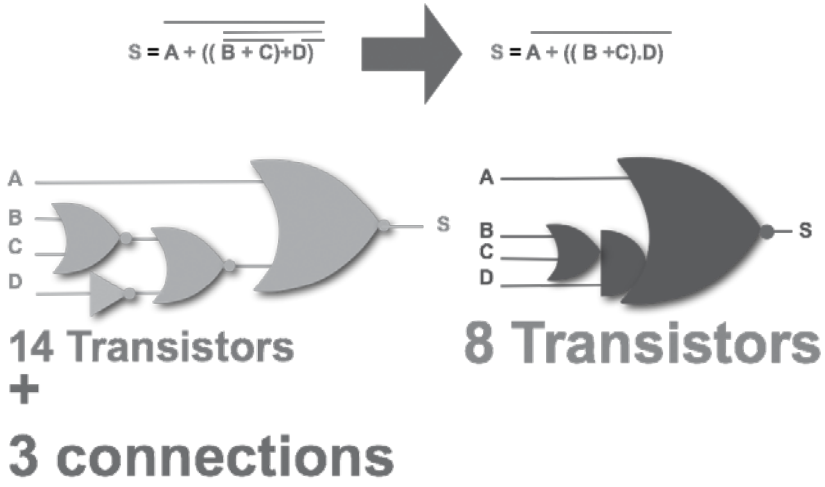


Figura 4 - A implementação da função em apenas uma porta lógica reduz também o número de conexões.

As bibliotecas de células comerciais possuem geralmente não mais do que 100 funções lógicas diferentes, sendo que tipicamente cada função é implementada com 3 versões de dimensionamento: uma versão para baixo consumo, outra para menor área e outra para menor atraso. Por outro lado, se limitarmos a geração de células em funções com 4 transistores em série, podemos chegar a 3503 funções lógicas diferentes como mostra a Tabela 1 [Detjens 1987].

Se limitarmos em 4 o número de transistores P em série e em 5 número de transistores N em série, chegamos a 28435 funções lógicas diferentes, ou seja, uma possibilidade de otimização da lógica, do que quando houver a disponibilidade de apenas as cerca de 100 funções lógicas de uma biblioteca de células típica.

TABELA 1 - Número de funções possíveis em função do número de transistores em série [Detjens 1987]

		Número de transistores em série PMOS				
		1	2	3	4	5
Número de transistores em série NMOS	1	1	2	3	4	5
	2	2	7	18	42	90
	3	3	18	87	396	1677
	4	4	42	396	3503	28435
	5	5	90	1677	28435	125803

Considerando que o consumo estático ganhou uma grande importância nas tecnologias CMOs recentes e que está relacionado ao número de transistores de um chip, entendemos que é tempo de efetuar esforços de otimização do número de transistores, de forma reduzir o consumo de um circuito integrado. Como observado no item anterior, o uso de biblioteca de células (metodologia Standard Cell) não permite otimizar o número de transistores, pois uma biblioteca de células padrão geralmente não contém mais de cerca de 100 funções lógicas.

A solução é contar com um método de síntese física que permita implementar qualquer rede de transistores. Em [Lubaszewski 1990] [Moraes 1994] [Moraes 1997] [Reis 1997] [Lazzari 2006] [Reis 2011] podem ser vistos alguns trabalhos visando a síntese automática de redes de transistores. Tendo uma ferramenta de EDA para síntese automática de rede de transistores, pode-se efetuar qualquer otimização lógica, pois sabe-se que a ferramenta de síntese do leiaute tem a capacidade de gerar o leiaute da função otimizada. Uma alternativa é fazer uso de uma ferramenta de síntese automática de leiaute, como o ASTRAN [Ziesemer 2007] [Ziesemer 2014], que implemente o leiaute de qualquer rede de transistores [Figura 5, Figura 6].

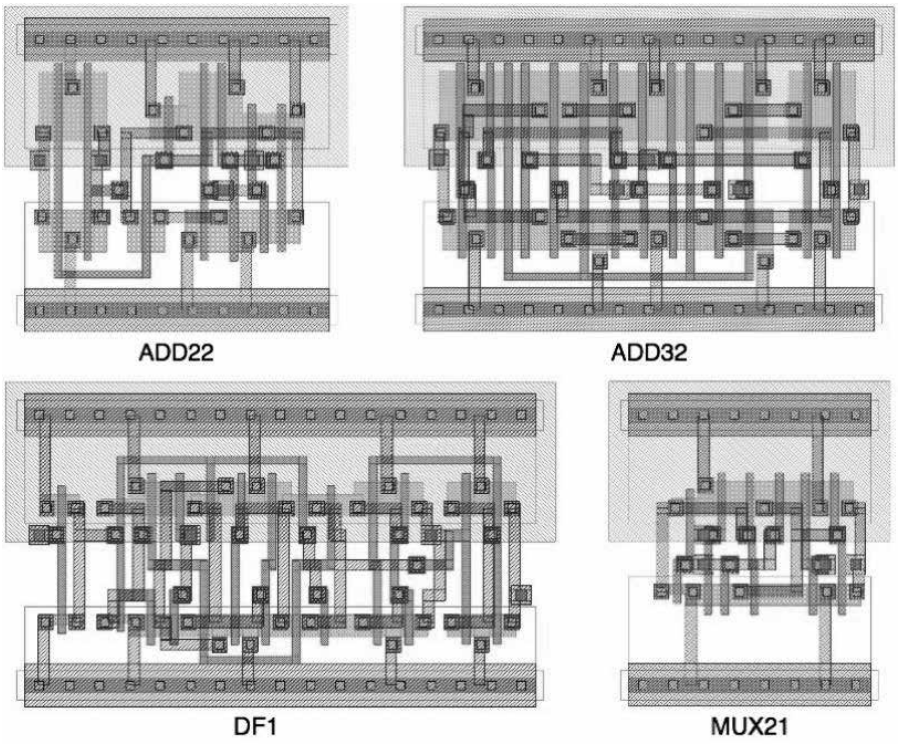


Figura 5: Leiaute de algumas redes de transistores geradas automaticamente pela ferramenta ASTRAN

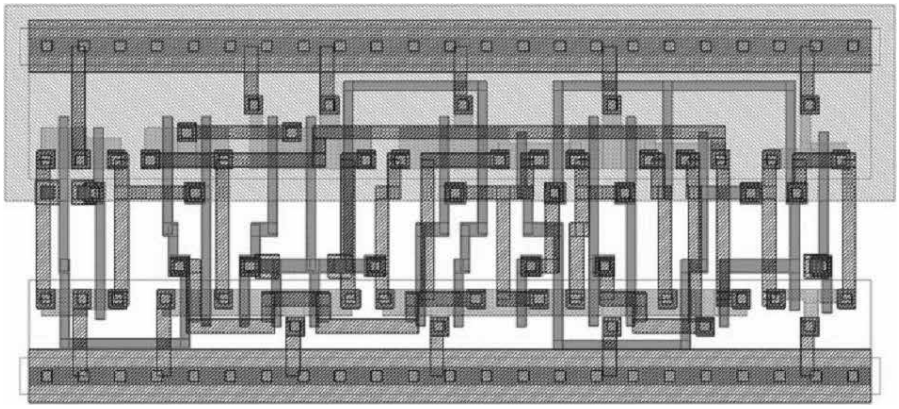


Figura 6: Leiaute de um Flip-Flop JK com 64 transistores gerado automaticamente pela ferramenta ASTRAN

4 TENDÊNCIAS EM EDA (ELECTRONIC DESIGN AUTOMATION)

A área de automação do projeto de chips tem evoluído constantemente ao longo do tempo de forma à atender aos novos desafios que a alta escala de transistores tem proporcionado. Inicialmente chamada de CAD (Computer Aided Design) para microeletrônica, hoje tem sido usado uma denominação mais precisa, ou seja, EDA (Electronic Design Automation). Sem o uso de ferramentas de EDA, fica impossível desenvolver o projeto de um circuito integrado no estado da arte, não só devido à complexidade dos mesmos, mas devido à necessidade de projetar um sistema integrado em um tempo viável.

O projeto de sistemas integrados complexos é desenvolvido atualmente através da especificação do mesmo usando uma descrição com o mais alto possível nível de abstração [Gajski 1988] [Reis 2009]. Esta descrição então é "convertida" para uma descrição mais detalhada, e portanto em um nível inferior de abstração. Este processo é chamado de síntese, ou seja, em um processo de síntese, descrições mais abstratas são transformadas (convertidas) em descrições mais detalhadas e assim sucessivamente até a obtenção da descrição do leiaute das máscaras que serão usadas no processo de fabricação. Portanto as ferramentas de EDA usadas neste processo são denominadas de ferramentas de síntese. Na Figura 7 pode ser observado um fluxo de síntese bastante simplificado descrevendo os principais níveis de abstração de um projeto de um circuito integrado.

Após a realização de uma etapa de síntese, é importante a realização de uma etapa de verificação, que confirme que a descrição obtida no processo de síntese é correspondente à descrição anterior. As ferramentas mais tradicionais de verificação são as ferramentas de simulação, mas devido à complexidade de muitos circuitos integrados

atuais, houve um desenvolvimento expressivo de ferramentas de verificação formal, que demonstram formalmente que a descrição de um sistema integrado está correto.

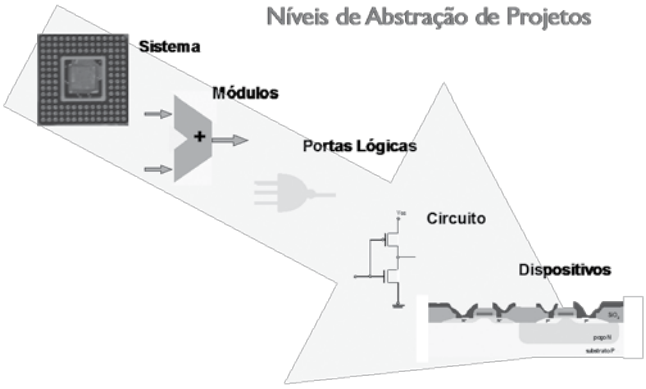


Figura 7 - Níveis de abstração no processo de síntese de sistemas integrados em um chip

Sobre cada uma das descrições em um nível de abstração determinado podem ser usadas ferramentas de EDA para a otimização da descrição. Estas ferramentas devem gerar uma descrição otimizada, usando a mesma linguagem da descrição inicial e o mesmo nível de abstração. Tanto no processo de síntese, como de otimização, é necessário o uso de ferramentas de estimativa. Ou seja, ferramentas de EDA que efetuam uma estimativa de área, número de componentes, consumo de energia, frequência de operação, etc... Estas estimativas são importantes para a tomada de decisão nas diferentes etapas de projeto de um sistema integrado. Por exemplo, devemos ter ferramentas de EDA para a estimativa de consumo de energia, em todos os níveis de abstração, de forma que em cada um dos níveis de síntese possam ser tomadas as decisões necessárias para que o consumo atenda às especificações iniciais. As ferramentas devem ser modificadas em função do desenvolvimento das novas tecnologias. Por exemplo, uma ferramenta para a estimativa de consumo em uma etapa de alto nível de abstração, considerava basicamente uma estimativa do número de transições no circuito integrado, nas diferentes situações de uso, para construir uma estimativa de consumo usando tecnologias CMOs tradicionais, pois o consumo de um sistema CMOs era devido basicamente ao consumo dinâmico (portanto, devido às transições do circuito de "1" lógico para zero e de zero para "1". Entretanto, como na tecnologias CMOs recentes o consumo estático cresceu significativamente devido ao crescimento das correntes de fuga, chegando em muitos casos a ser superior ao consumo estático, as ferramentas de EDA para estimativa de consumo em alto nível de abstração devem ser modificadas. Ou seja, devem considerar que o consumo estático está relacionado ao número de transistores e portanto é necessário uma estimativa do número de transistores que o circuito integrado deverá

ter. Claro, que neste processo deve também ser considerado as diferentes técnicas para redução de consumo que devem ser aplicadas no projeto do sistema integrado em desenvolvimento.

Em parte devido ao fato de que o custo de transistores é relativamente barato, mas principalmente devido ao fato de que em tecnologias tradicionais o consumo era essencialmente o consumo dinâmico, não houve até agora muita preocupação em reduzir o número de transistores de um chip. Entretanto, como nas tecnologias CMOs recentes o consumo estático é tão ou mais expressivo, é importante a obtenção de novas metodologias de síntese física que tenham a preocupação de reduzir o número de transistores, como uma forma de reduzir consumo estático, assim como reduzir atraso (devido à redução do comprimento médio das conexões) [Reis 2011]. A redução do número de transistores provoca naturalmente redução de área, o que provoca a redução do comprimento médio das conexões, que por sua vez reduz o tempo de propagação dos sinais.

Como a área dos chips tem aumentado continuamente e como as dimensões dos transistores tem diminuido também continuamente, a propagação de um sinal pelas conexões longas que ligam circuitos em posições extremas no chip, pode demorar alguns cliclos de relógio. Para tratar deste problema, a implantação da rede de relógio tem um tratamento especial, utilizando técnicas que diminuam ou eliminem os problemas de atraso na propagação dos sinais da rede de relógio [Guthaus 2013].

Outra tendência em EDA é o desenvolvimento de ferramentas eficientes para o dimensionamento de transistores e definição de Vth destes transistores, como uma forma de contribuir para a redução de consumo, mas mantendo o desempenho desejado [Fang 1995] [Posser 2012] [Flach 2013] [Flach 2014].

A geração de chips para uso em sistemas críticas demanda observar diversos cuidados na realização da síntese física dos circuitos integrados, de forma que tenham um maior tempo de vida e uma maior confiabilidade. Por exemplo, a geração do leiuate deve observar técnicas que reduzam ou eliminem os mecanismos de eletromigração [Posser 2014]. É importante também o uso de técnicas que permitam tratar dos problemas de dissipação térmica, especialmente em ciurcuitos 3D [Santos 2014].

5. CONCLUSÕES

Por muito tempo, no projeto de sistemas computacionais e eletrônicos, não houve uma preocupação muito grande com a redução do consumo dos sistemas integrados. Atualmente, existem dois grandes motivos para uma grande preocupação com a redução de consumo. Um é a alta capacidade de integração, que permite colocar milhões de transistores em um chip, com uma densidade de transistores por área cada vez maior. Isto provoca não só um grande consumo, como uma alta densidade de consumo de

energia, gerando "pontos quentes" no chip que podem provocar falhas durante o seu funcionamento. Outro motivo é que as nanotecnologias CMOs atuais apresentam um alto consumo estático, o que no passado era desprezível, ou seja, o circuito consome, mesmo quando não existem transições de sinal. Este consumo estático está relacionado com o número de componentes. Portanto, a otimização do número de componentes de um sistema integrado é uma área que ganha uma relevância expressiva.

Na área médica, existem mais e mais aplicações que fazem uso da eletrônica e da computação embarcada em um chip. Especialmente os chips a serem implantados em seres humanos, devem ter ultra baixo consumo, assim como usarem técnicas de tolerância à falhas.

Portanto, a geração de circuitos integrados com ultra baixo consumo e com altas taxas de confiabilidade é certamente um dos Grandes Desafios da Computação para as próximas décadas.

6. REFERÊNCIAS

[SBC 2006] SBC (2006) Brazilian Computer Society. Grand Challenges in Computer Science Research in Brazil 2006-2016, 25pgs. In: http://www.sistemas.sbc.org.br. Último acesso em 22 de março de 2010.

[BAMPI 2009] BAMPI, S., SUSIN, A., REIS, R., Systems Architectural Challenges for Transitional and Compatible to CMOS Technologies in Giga-Scale Hardware Integration, SEMISH 2009, Anais do 36º Seminário Integrado de Software e Hardware, Bento Gonçalves, 21 a 22 de Julho de 2009, p. 281-292, ISSN: 2175-2761.

[ITRS 2009], International Roadmap Committee. "The International Technology Roadmap for Semiconductors - 2009". In http://www.itrs.net/home.html. Último acesso 26 de março de 2010.

[Economist 2010] Revista The Economist, 6 de Setembro de 2010

[SIA 2011] - Semiconductor Industry Association, The Technology Roadmap for Semiconductors 2011, http://www.itrs.org/

[REIS 2009] REIS, Ricardo e Cols.. Concepção de Circuitos Integrados, 2ª Edição. Série Livros Didáticos do Instituto de Informática, Editora Bookmann, Porto Alegre, 2009, 258 páginas. ISBN 9788577803477

[Guthaus 2013] GUTHAUS, M., WILKE, G., REIS, R., Revisiting Automated Physical Synthesis of High-Performance Clock Networks, ACM TODAES - ACM Transactions on Design Automation of Electronic Systems, Vol. 18, Issue 2, DOI: 10.1145/2442087.2442102, ISSN: 1084-4309, EISSN:1557-7309, March 2013.

[Brayton 1984] BRAYTON, R. K. et al. Logic Minimization Algorithms for VLSI Synthesis. Kluwer, 1984.

[Jacobi 1996] JACOBI, Ricardo. Síntese de Circuitos Lógicos Combinacionais. 10ª Escola de Computação, Campinas, Instituto de Computação, UNICAMP, 1996,169 páginas.

[Detjens 1987] DETJENS E. et al. Technology Mapping in MIS, IEEE ICCAD, Proceedings..., pp. 116-119, 1987.

[Lubaszewski 1990] LUBASZEWSKI, Marcelo S. Geração Automática de Lógica Aleatória Utilizando a Metodologia TRANCA. Porto Alegre: CPGCC da UFRGS, 1990. 232p. Dissertação de Mestrado.

[Lazzari 2006] LAZZARI, C., SANTOS, C., REIS, R., A New Transistor-Level Layout Generation Strategy for Static CMOS Circuits, 13th IEEE International Conference on Electronics, Circuits and Systems – ICECS2006, Nice, France, December 10 - 13, 2006, p. 660-663, ISBN: 1-4244-0395-2, DOI 10.1109/ICECS.2006.379875.

[Moraes 1994] MORAES, Fernando. Synthèse Topologique de Macro-Cellules en Technologie CMOS. Montpellier (França): Université Montpellier II, 1994. Tese de Doutorado.

[Moraes 1997] MORAES, F.; REIS, R.; LIMA F. An Efficient Layout Style for Three-Metal CMOS Macro-Cells. In: IFIP International Symposium on Very Large Scale Integration, 1997 Gramado-Brazil Proceedings... IFIP, 1997. p 415-426.

[Reis 1997] REIS, André; REIS, Ricardo; ROBERT, Michel; AUVERGNE, Daniel, Library Free Technology Mapping. In: VLSI: Integrated Systems on Silicon, Chapman & Hall, London, 1997, pg. 303-314, ISBN 0-412-82370-5

[Reis 2011] REIS, R., Design Automation of Transistor Networks, a New Challenge. IEEE International Symposium on Circuits and Systems, ISCAS2011, Rio de Janeiro, Brasil, May 15-19, 2011. IEEE Press. p. 2485-2488, ISBN: 978-1-4244-9472-9. DOI 10.1109/ISCAS.2011.5938108

[Ziesemer 2007] ZIESEMER, A.; LAZZARI, C., REIS, R., Transistor Level Automatic Layout Generator for non-Complementary CMOS Cells, In: IFIP/CEDA VLSI-SoC2007, International Conference on Very Large Scale Integration, Atlanta, USA, October 15-17, 2007. pp. 116-121, ISBN: 978-1-4244-1710-0, DOI 10.1109/VLSISOC.2007.4402483

[Ziesemer 2014] ZIESEMER, A., REIS, R., Simultaneous Two-Dimensional Cell Layout Compaction Using MILP with ASTRAN, ISVLSI2014 - IEEE Computer Society Annual Symposium on VLSI, July 9-11, 2014, Tampa, USA, p. 350-355, ISBN: 978-1-4799-3765-3, DOI 10.1109/ISVLSI.2014.79

[Gajski 1988] GAJSKI, Daniel D. Silicon Compilation, Adison-Wesley Publishing Company, 1988, 450p..

REIS, R., Power Consumption & Reliability in NanoCMOS, IEEE NANO, 11th International Conference on Nanotechnology, Portland, USA, August 15-19, 2011 (invited talk). p.711-714. ISBN 978-1-4577-1515-0, DOI: 10.1109/NANO.2011.6144656

[Fang 1995] FANG, C.-L.; JONE, W.-B. Timing Optimization by Gate Resizing and Critical Path Identification. IEEE Transactions on CAD, v.14, n.2, p.201-217., February 1995.

[Posser 2012] POSSER, G., FLACH, G., WILKE, G., REIS, R., Gate Sizing using Geometric Programming, IN: Analog Integrated Circuits and Signal Processing, Volume 73, Number 3, 831-840, December 2012, Springer, ISSN 0925-1030, DOI: 10.1007/s10470-012-9943-3.

[Flach 2013] FLACH, G., REIMANN, T., POSSER, G., JOHANN, G., REIS, R., Simultaneous Gate Sizing and Vth Assignment using Lagrangian Relaxation and Delay Sensitivities, ISVLSI2013. IEEE Computer Society Annual Symposium on VLSI, Natal, Brazil, August 5-7, 2013.

[Flach 2014] FLACH, G., REIMANN, T., POSSER, G., JOHANN, G., REIS, R., An Effective Method for Simultaneous Gate Sizing and Vth Assignment using Lagrangian Relaxation, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 33, Issue 4, April 2014, p. 546-557, DOI: 10.1109/TCAD.2014.2305847, ISSN 0278-0070.

[Posser 2014] POSSER, G., MISHRA, V., JAIN, P., REIS, R., SAPATNEKAR, S., A Systematic Approach for Analyzing and Optimizing Cell-Internal Signal Electromigration, ICCAD 2014 – 33rd IEEE/ACM International Conference on Computer-Aided Design, November 3-6, San Jose, USA. p. 486-491, ISBN: 978-1-4799-6278-5

[Santos 2014] SANTOS, C., VIVET, P., COLONNA, J.-P., COUDRAIN, P., REIS, R., Thermal Performance of 3D ICs: Analysis and Alternatives, 3DIC – 4th IEEE EEE International Conference on 3D System Integration, December 1-3, 2014, Cork, Ireland.

ENRIQUECIMENTO SEMÂNTICO, ANÁLISE E MINERAÇÃO DE DADOS SOBRE MOVIMENTO COM ONTOLOGIAS E DADOS LIGADOS

Renato Fileto[1] José A. F. de Macêdo[2] Vania Bogorny[1]
Chiara Renso[3] Alessandra Raffaetà[4] Nikos Pelekis[5] Yannis Theodoridis[5]

Tema: GESTÃO DA INFORMAÇÃO EM GRANDES VOLUMES DE DADOS MULTIMÍDIA DISTRIBUÍDOS

Business Intelligence (Sistemas analíticos e de predição, *Big Data*);

Mineração de dados;

Fusão de dados;

Integração de sistemas, dados (desestruturados e heterogêneos) e informações;

Interoperabilidade semântica (ontologias);

Recuperação de Informação, Busca semântica;

Redes Sociais.

Domínios de interesse: Outros (transportes, mobilidade, logística, segurança pública, planejamento de infra-estrutura e serviços).

Palavras-chaves: trajetórias de objetos móveis, postagens em mídias sociais, ontologias, dados ligados abertos (*LOD*), *data warehousing*, mineração de dados.

1Depto. de Informática e Estatística, Univ. Federal de Santa Catarina (INE/UFSC) Caixa Postal 476, 88.040-900, Florianópolis-SC, BRASIL
{fileto|bogorny}@inf.ufsc.br

2Departamento de Computação, Universidade Federal do Ceará (DC/UFC) Campus do Pici, Bloco 910,60.455-760, Fortaleza-CE, BRASIL
jose.macedo@lia.ufc.br

[3]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (KDDLab/ISTI/CNR) via G. Moruzzi 1, 56124, Pisa, ITALIA
{chiara.renso|mirco.nanni}@isti.cnr.it

[4]Dipartimento di Informatica, Universita` Ca'Foscari di Venezia (DAIS/UNIVE) Via Torino 155, 30172, Mestre (VE), ITALIA
{raffaeta|roncato}@dsi.unive.it

[5]Information Management Laboratory at University of Piraeus (InfoLab/UPRC), 80 Karaoli & Dimitriou Str., 18534, Piraeus, GREECE
{yannis|nikos}@unipi.gr

Abstract. *The widespread use of sensors and information systems, frequently via mobile devices, has created an abundance of data about movement, such as moving objects' trajectories and social media user's posts. These data can be valuable in several applications, whose realization requires the development of new methods for analyzing movement data. The recent progress in this area mostly considers spatiotemporal data. However, other data (e.g., tags, comments) could also help to better understand movements (e.g., visited places and events, reasons for stops and moves). In addition, movement data must be enriched with well-defined semantics, processable by machines, to enable the precise identification of specific objects, phenomena, actions (e.g., buy or watch the movie Rio), and concepts (e.g., movie, city, state) that are relevant for information extraction and its proper understanding. This work proposes the development of methods for: fusing trajectories with social media posts based on their spatiotemporal proximity, to produce detailed data about the movements annotated with the posts contents; and (ii) connecting those freely annotated movement data to linked data, to support sophisticated queries and analysis. This proposal has potential to produce relevant scientific contributions, as demonstrated by preliminary results. In addition, it is expected to have considerable impact on applications of interest in several domains, and future research in the area of knowledge discovery in movement databases.*

Resumo. *O uso extensivo de sensores e sistemas de informação, muitas vezes através de dispositivos móveis, criou uma abundância de dados sobre o movimento, como trajetórias de objetos móveis e postagens em mídias sociais. Estes dados podem ser valiosos em várias aplicações, cuja realização requer o desenvolvimento de novos métodos para analisar dados sobre movimento. O recente progresso nesta área considera principalmente dados espaço-temporais. No entanto, outros dados (e.g., rótulos, comentários) também podem ajudar a compreender melhor os movimentos (e.g., lugares e eventos visitados, razões para paradas e movimentos). Além disso, dados sobre movimento devem ser enriquecidos com semântica bem definida, processável por máquinas, para per- mitir a identificação precisa de objetos, fenômenos, ações (e.g., comprar ou assistir o filme Rio) e conceitos (e.g., filme, cidade, estado) que são relevantes para a extração de informação e a sua compreensão adequada. Este trabalho propõe o desenvolvimento de métodos para: (i) fundir espaço-temporalmente trajetórias com postagens em mídias sociais, de modo a produzir dados detalhados sobre movimentos, anotados com o conteúdo das postagens; e (ii) conectar esses dados sobre movimento livremente anotados a dados ligados, para possibilitar consultas e análises sofisticadas. Esta proposta tem potencial para produzir contribuições científicas relevantes, como demonstrado por resultados preliminares. Além disso, espera-se que ela tenha um impacto considerável sobre aplicações de interesse em diversos domínios, e futuras pesquisas na área de descoberta de conhecimento em dados sobre movimento.*

INTRODUÇÃO E MOTIVAÇÃO

A popularização de dispositivos móveis, junto a tecnologias de posicionamento e sensoreamento (e.g., GPS, GSM, RFID, câmeras), e a disseminação do uso de sistemas de informação na Web (por exemplo, mídias sociais) tem criado uma enorme abundância de dados sobre a movimentação de objetos e seres (veículos, pessoas, etc.). Diversas aplicações podem se beneficiar da análise desses dados, em domínios tão diversos como logística, transportes, segurança pública e publicidade. No entanto, para realizar potenciais aplicações é necessário desenvolver métodos apropriados para extrair informação útil de grandes quantidades de dados sobre movimento.

Nesta proposta, chamamos *dados sobre movimento* qualquer coleção de posições espaço-temporais amostradas ou inferidas de objetos móveis. Elas podem ser capturadas por sensores e/ou sistemas de informação. Cada posição do objeto móvel pode ser representada por coordenadas geográficas e um indicador do momento em que o objeto esteve em tal posição. *Dados sobre movimento livremente anotados* têm texto associado a (algumas) amostras de posição [May and Fileto, 2014]. Esta definição engloba, entre outras coisas, trajetórias de objetos móveis com segmentos (sub-trajetórias) anotados, sequências de postagens de usuários em mídias sociais e fusões espaço-temporais das mesmas [Nabo et al., 2014].

Uma *Trajetória* é uma sequência temporalmente ordenada de posições espaço-temporais ocupadas por um objeto em movimento. Trajetórias são coletadas mediante o emprego de sensores e aplicações para fins específicos (rastreamento, navegação, etc.). Hoje em dia, é possível obter trajetórias precisas, usando sensores modernos e taxas de amostragem finas (por exemplo, a cada segundo, a cada 3 metros). No entanto, é difícil conseguir grandes volumes de trajetórias anotadas, porque anotar é tarefa laboriosa. Chamamos de *trilha de usuário* uma sequência temporalmente ordenada de postagens do mesmo usuário em um mesmo sistema (por exemplo, Twitter, Facebook, FourSquare). Diferentemente de trajetórias, trilhas de usuários em mídias sociais são geralmente imprecisas e esparsas no espaço e no tempo, devido às limitações que podem ser impostas no acesso a posições dos usuários e à natureza assíncrona das postagens dos usuários. No entanto, as postagens em mídias sociais geralmente são ricas em conteúdos (por exemplo, texto, *hashtags*, nomes de locais e eventos). Alguns desses conteúdos podem servir como anotações e auxiliar a explicar os movimentos, por meio de indicações de aspectos semânticos relevantes (e.g., objetivos, locais e eventos visitados).

Recentemente, tem havido progressos significativos em métodos para lidar com dados sobre movimento [Spaccapietra et al., 2008, Nanni et al., 2010, Giannotti et al., 2011, Furletti et al., 2013, Pelekis and Theodoridis, 2014]. Grande parte deste progresso se refere a manipulação e análise de dados espaço-temporais. Entretanto, a comunidade científica reconhece que questões semânticas devem ser abordadas para melhor com-

preender e explorar dados sobre movimento [Spaccapietra and Parent, 2011, Parent et al., 2013, Yan et al., 2013, Renso et al., 2013, Furtado et al., 2013, Fileto et al., 2013, Bogorny et al., 2014, Pelekis and Theodoridis, 2014]. Por exemplo, considere as seguintes questões propostas em artigos a respeito de análise de dados sobre movimento:

Q1: *"Selecione as trajetórias que tenham pelo menos uma parada relacionada com algum evento esportivo."* [Fileto et al., 2013]

Q2: *"Selecione as trajetórias que tenham pelo menos uma parada localizada dentro de uma dada distância do local turístico chamado Corcovado na cidade do Rio de Janeiro."* [Fileto et al., 2013]

Q3: *"Qual a distância média percorrida por pessoas usando transporte público para visitar ao menos uma atração cultural?"* [Wagner et al., 2013]

Q4: *"Quais as trajetórias de pessoas que trabalham pelo menos duas unidades de tempo, e então, depois de no máximo quatro unidades de tempo, param em casa para jantar?"* [Simões et al., 2012]

Note que estas consultas envolvem aspectos semânticos do movimento, referindo-se a conceitos como *local turístico*, *pessoas*, *transporte público*, *atração cultural*, *trabalho*, *casa* e *jantar*. Algumas delas também se referem a objetos específicos como o local turístico chamado *Corcovado* e a cidade chamada *Rio de Janeiro*. Consequentemente, não é possível resolvê-las com confiança considerando somente as coordenadas espaço-temporais. Antes de resolver essas consultas, os dados sobre movimento a serem consultados precisam ser enriquecidos semanticamente com anotações que permitam a identificação precisa das classes e instâncias relevantes em determinadas posições espaço-temporais ou segmentos do movimento.

Nosso trabalho deriva do projeto SEEK (SEmantic Enrichment of trajectory Knowledge discovery)[1] e tem por objetivo explorar técnicas de mineração de dados, ontologias e dados ligados[2] na manipulação, no enriquecimento semântico e na análise de dados sobre movimento. Temos desenvolvido diversos métodos e algoritmos para extração de episódios (trechos satisfazendo certos predicados) de sequências de dados sobre movimento, enriquecimento semântico dos dados sobre movimento e dos episódios deles extraídos e detecção de padrões de movimento e comportamento de objetos móveis. Atualmente, estamos desenvolvendo métodos e algoritmos para fundir trajetórias com trilhas de mídias sociais produzindo trajetórias livremente anotadas, e para conectar dados sobre movimento livremente anotados com dados ligados produzindo anotações com semântica melhor definida do que texto livre. Para fazer isso, temos empregado técnicas de fusão de informações, compatibilidade espaço-temporal, vários tipos de semelhança léxica, junções por similaridade e casamento de entidades (*entity linking*), entre outras possibilidades. A semântica bem definida de classes e instâncias em da-

dos sobre movimento anotados com dados ligados viabiliza a execução de consultas tais como expressões em geoSPARQL [Battle and Kolas, 2012] para **Q1** e **Q2**, apresentadas em [Fileto et al., 2013]. Além disso, estamos começando a utilizar os dados sobre movimento semanticamente enriquecidos com ontologias e dados ligados, produzidos por alguns métodos e algoritmos já desenvolvidos, em de análise de informações em armazéns de dados (*data warehouses*) e novas tarefas de mineração de dados.

O restante desta proposta está assim organizado. A Seção 2 descreve os objetivos das nossas pesquisas. A Seção 3 descreve o plano inicial para alcançar tais objetivos. A Seção 4 discute alguns trabalhos relacionados. Finalmente, a Seção 5 delineia as conclusões e os impactos esperados das nossas pesquisas.

OBJETIVOS

O objetivo geral das nossas pesquisas é desenvolver métodos para enriquecer semanticamente dados sobre movimento livremente anotados com ontologias e dados ligados, e então utilizar os dados resultantes para alimentar processos de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases – KDD*), baseados em armazéns de dados e novas técnicas de mineração de dados que explorem a semântica bem definida utilizada para enriquecer os dados sobre movimento. Para alcançar este objetivo os seguintes problemas têm sido abordados:

obter ontologias e dados ligados apropriados para o enriquecimento semântico de dados sobre movimento;

• selecionar conceitos e propriedades das ontologias selecionadas que tenham maior relevância no domínio de aplicação;

• coletar dados sobre movimento com anotações na forma de texto livre associadas a (algumas) amostras de posição espaço-temporal;

• conectar dados sobre movimento livremente anotados a dados ligados, para obter anotações com semântica mais específica e precisa que anotações textuais livres;

• avaliar a complexidade computacional, a velocidade de execução e a qualidade dos resultados obtidos pelos métodos propostos;

• investigar a adoção de técnicas de processamento paralelo e *big data*, para lidar com o volume, a velocidade de coleta e a variedade, dos dados sobre movimento.

[1]http://www.seek-project.eu
[2]http://www.w3.org/standards/semanticweb/data

Obviamente, é difícil coletar grandes quantidades de trajetórias anotadas para ajustar e testar métodos de enriquecimento semântico. Planejamos contornar este problema explorando a riqueza de dados sobre movimento livremente anotados disponíveis atualmente em mídias sociais. A trilhas de usuários em tais sistemas são assíncronas, espacial e temporalmente esparsas e menos precisas do que trajetórias colhidas por aplicações voltadas para monitoramento de movimento utilizando sensores. No entanto, as postagens em mídias sociais geralmente têm abundância de conteúdos associados que podem ser úteis para entender melhor o movimento (locais e/ou eventos visitados, razões das movimentações, etc.)

Os mesmos dispositivos (por exemplo, telefones celulares) podem coletar trajetórias dos seus portadores e concomitantemente serem usados para fazer postagens em mídias sociais. Mesmo quando trajetórias e trilhas são colhidas por diferentes dispositivos, elas (ou segmentos das mesmas) podem se referir à mesma pessoa ou a pessoas com comportamentos análogos, quando elas coincidem ou estão próximas no espaço e no tempo (por exemplo, uma pessoa fazendo postagens ao ser levada em um veículo monitorado, pessoas participando de uma parada ou demonstração). Assim, nós pretendemos fundir trajetórias com trilhas de usuários em redes sociais, para produzir trajetórias anotadas com o conteúdo de postagens, e então conectar as trajetórias anotadas resultantes a dados ligados.

Os objetivos específicos das nossas pesquisas são:

Desenvolver métodos para **fundir trajetórias com trilhas de mídias sociais**, de modo a produzir trajetórias anotadas com o conteúdo das postagens. O processo de fusão basear-se-á em compatibilidade espaço-temporal, ou seja, segmentos de trajetória (por exemplo, paradas) serão fundidos com postagens que ocorrem em torno do mesmo tempo e lugar. O trabalho neste tema foi iniciado por alunos brasileiros e gregos, supervisionados por pesquisadores brasileiros, gregos e italianos. Um relato de métodos e experimentos preliminares pode se encontrado em [Nabo et al., 2014].

Desenvolver métodos para **enriquecer semanticamente dados sobre movimento**. Pretendemos fazer isso principalmente através da ligação de segmentos de movimento livremente anotados a classes e instâncias descritas em ontologias e coleções de dados ligados. Em primeiro lugar, pretendemos filtrar conexões candidatas com base na compatibilidade espaço-temporal e na proximidade léxica entre os conteúdos textuais associados aos segmentos de movimento e a recursos de dados ligados (por exemplo, *tags* de postagens em mídias sociais e rótulos de recursos de dados ligados). Em seguida, pretendemos explorar expansão semântica e informação de contexto, entre outras possibilidades, para melhorar a cobertura e a precisão das conexões obtidas. Os resultados preliminares de nossa pesquisa neste tema podem ser encontrados em [Fileto et al., 2013] e [May and Fileto, 2014].

Avaliar a eficiência e a eficácia dos métodos propostos para enriquecer semanticamente dados sobre movimento através de análise teórica e experimentos. Nossa intenção é medir em experimentos a velocidade de execução dos métodos, bem como a cobertura e a precisão dos resultados obtidos.

Efetuar consultas e desenvolver novas técnicas de mineração de dados sobre as coleções de dados sobre movimento semanticamente enriquecidos, a fim de explorar as anotações dos movimentos com recursos de ontologias e dados ligados obtido nos níveis extensional (instâncias e suas relações) e intencional (conceitos e suas relações). Esses trabalhos também vão ajudar a avaliar os dados gerados pelos métodos propostos no suporte a novas técnicas de mineração de dados e a consultas como as apresentadas na Seção 1 e vários trabalhos relacionados [Simões et al., 2012, Fileto et al., 2013, Wagner et al., 2013, Leonardi et al., 2014, Zhang et al., 2014].

PLANO PRELIMINAR DE DESENVOLVIMENTO

Trajetórias para realizar experimentos estão sendo fornecidas pelos parceiros do projeto e demais colaboradores. Postagens de usuários em mídias sociais e dados ligados estão sendo obtidos de coleções disponíveis na Web através de respectivas APIs. Alguns dados históricos de mídias sociais já foram coletados em trabalhos anteriores. Voluntários usando ferramentas para coleta de trajetórias e acessando redes sociais através de seus próprios telefones celulares começam a contribuir com dados sobre movimento adicionais e anotações feitas manualmente, através de iniciativas como TagMyDay[3]. Alguns dos dados assim coletados e anotados servirão como regra ouro para analisar a precisão e a cobertura dos métodos propostos.

Planejamento de atividades

- Seguem as atividades planejadas para o projeto proposto. A Tabela 1 apresenta o cronograma proposto para executar tais tarefas em 36 meses.

- Pesquisa bibliográfica e análise de sistemas relacionados aos problemas tratados no projeto.

- Coleta, preprocessamento e catalogação de trajetórias de objetos móveis, trilhas de usuários em redes sociais e coleções de dados ligados para experimentos.

- Elicitação dos requisitos dos métodos a serem desenvolvidos.

- Projeto, implementação e validação de métodos para fundir trajetórias com trilhas.

[3]http://tagmyday.isti.cnr.it/

• Desenvolvimento, implementação e teste do processo para converter dados geográficos livremente anotados em dados ligados.

• Projeto, implementação e validação de métodos para conectar dados sobre movimento livremente anotados com dados ligados.

• Experimentos de execução de consultas em coleções de dados sobre movimento enriquecidas semanticamente com dados ligados obtidas da aplicação de processos e métodos produzidos nas tarefas anteriores aos dados coletados para experimentos.

• Disseminação dos resultados em relatórios técnicos, artigos, monografias e documentação de software e bases de dados produzidas.

Atividade	Meses											
	01-03	04-06	07-09	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	32-36
1	√	√			√				√			
2	√	√	√	√	√	√	√	√	√	√		
3	√	√	√									
4		√	√	√	√	√	√	√				
5			√	√	√	√	√	√	√			
6			√	√	√	√	√	√	√	√	√	
7					√	√	√	√	√	√	√	√
8				√	√	√	√	√	√	√	√	√

Tabela 1. Cronograma

TRABALHOS RELACIONADOS

Os pesquisadores envolvidos nesta proposta têm contribuído com trabalhos relevantes em diversos temas relacionados à análise e à semântica de dados sobre movimento, incluindo:

• modelos, métodos e ferramentas para tratamento de dados espaço-temporais e mineração de dados sobre movimento [Pelekis et al., 2008, Palma et al., 2008, Rocha et al., 2010, Nanni et al., 2010, Giannotti et al., 2011, Bogorny et al., 2011, Parent et al., 2013, de Aquino et al., 2013, Moreno et al., 2014, Bogorny et al., 2014, Pelekis and Theodoridis, 2014];

• métodos para detectar comportamentos, padrões e outros tipos de fenômenos mediante o processamento de dados sobre movimento [Brilhante et al., 2012, Alvares et al., 2011, Furletti et al., 2013, Furtado et al., 2013, Renso et al., 2013, Gabrielli et al., 2013, Barbosa et al., 2013];

• progressos em consultas a dados sobre movimento, extensões espaciais e temporais para armazéns de dados e armazéns de dados sobre movimento [Deggau et al., 2010, Filho et al., 2010, Arboleda et al., 2010, Wachowicz et al., 2011, Simões et al., 2012, Pelekis et al., 2013, Wagner et al., 2013, Leonardi et al., 2014];

• métodos e ferramentas para anotação de dados e buscas semânticas [Fileto et al., 2005, D'Agostini et al., 2008, D'Agostini and Fileto, 2009, da Rocha et al., 2009, Rinzivillo et al., 2013, Nabo et al., 2014]; and

• uso de ontologias e dados ligados para o enriquecimento semântico e a análise de dados sobre movimento [Manco et al., 2008, Baglioni et al., 2008, Baglioni et al., 2009, Fileto et al., 2013, May and Fileto, 2014].

A idéia de enriquecer semanticamente dados sobre movimento com dados ligados foi introduzida em [Fileto et al., 2013], que propõe construções ontológicas e um processo semi-automático para tal enriquecimento. Tal trabalho também ilustra os benefícios da semântica bem definida por trás de dados ligados, quando estes são usados em anotações de segmentos de dados sobre movimento, para suportar a execução de consultas envolvendo aspectos semânticos do movimento e do ambiente onde ele ocorre. Porém, o desenvolvimento e a avaliação de métodos eficientes e eficazes para enriquecer dados sobre movimento com ontologias e dados ligados ainda é um desafio de pesquisa aberto.

A fusão de trajetórias com trilhas (sequências de postagens) de usuários em mídias sociais visando produzir trajetórias anotadas é inicialmente tratada em [Nabo et al., 2014]. As trajetórias são segmentadas e estruturadas como sequências temporalmente ordenadas de paradas e movimentações, utilizando técnicas baseadas em agrupamento, tais como CB-SMOT [Palma et al., 2008] e DB-SMOT[Rocha et al., 2010]. Em seguida, as posições espaço-temporais de ambas, as trajetórias estruturadas e as trilhas , são indexadas para suportar algoritmos eficientes que usam junção por proximidade para fundir as trajetórias estruturadas com trilhas.

O trabalho em curso descrito em [May and Fileto, 2014] é apenas um esforço inicial para conectar dados sobre movimento livremente anotados a dados ligados. Tal trabalho emprega Soft-TF-IDF [Moreau et al., 2008] para calcular a similaridade textual entre dados textuais associados aos segmentos de movimento e a recursos de dados ligados, com base em alguma métrica de distância entre palavras, tais como a distância de edição de Levenshtein ou Jaro Winkler [Navarro, 2001, Cohen et al., 2003]. Primeiramente, utiliza-se um método de acesso espacial para filtrar os recursos que estão suficientemente perto (dentro de um raio dado) de cada segmento de movimento anotado. Em seguida, filtram-se os recursos que apresentam maior similaridade textual Soft-TF-IDF, dentre aqueles espacialmente próximos do segmento de movimento.

Experimentos preliminares relatados em [Fileto et al., 2013] e [May and Fileto, 2014] usam esse método para ligar postagens de fotos feitas por um mesmo usuário no Flickr com recursos das coleções DBpedia e LinkedGeoData, com base na proximidade espacial e em similaridade textual entre as etiquetas (*tags*) das postagens e os rótulos dos recursos. Apesar da simplicidade do método e da baixa probabilidade de uma postagem no Flickr referir-se a um lugar descrito na DBPedia, os resultados incluem algumas ligações confirmadas por inspeção fotográfica e textual. Algumas dessas ligações, principalmente em áreas com alta densidade de locais de interesse, foram claramente possibilitadas pela similaridade textual, uma vez que as coordenadas espaciais não têm precisão suficiente para permitir a ligação correta. Técnicas mais sofisticadas podem ser usadas para auxiliar a conectar dados sobre movimento a dados ligados de maneira apropriada e eficientemente, uma vez que este problema está relacionado a consultas por palavras-chave e concomitantemente espaciais ou espaço-temporais [De Felipe et al., 2008, Cong et al., 2012, Chen et al., 2013, Zhang et al., 2013], casamento de entidades (*entity linking*) [Ceccarelli et al., 2013, Shen et al., 2013] e junções por similaridade [Bouros et al., 2012, Liu et al., 2012].

CONCLUSOES

Esta proposta tem como objetivo produzir novos métodos para enriquecer semanticamente dados sobre movimento (trajetórias, trilhas e fusões dos mesmos) com dados ligados. Nossa estratégia consiste em primeiro produzir dados sobre movimento livremente anotados, mediante a fusão de trajetórias com trilhas de usuários em redes sociais, e então conectar os segmentos anotados de dados sobre movimento a dados ligados. Até onde vai nosso conhecimento, tal proposta é original e inovadora. Esta agenda de pesquisa tem grande potencial para contribuir no enriquecimento semântico e na análise de dados sobre movimento, como tem sido demonstrado por resultados experimentais preliminares.

Os resultados dos trabalhos sendo conduzidos devem ter impacto tanto em aplicações quanto em pesquisas futuras na análise e extração de conhecimento dos dados sobre movimento. Consultas suportadas por dados sobre movimento semanticamente enriquecidos com ontologias e dados ligados têm aplicações em vários domínios. Os grupos de pesquisa envolvidos neste proposta têm convênios com empresas e instituições de diversos ramos de atividade, tais como telecomunicações, seguros, logística e serviços públicos. Algumas delas têm cedido dados reais para as nossas pesquisas e forte interesse em transformar os resultados da nossa pesquisa em soluções e produtos para utilização em larga escala. Também temos coletado grandes volumes de dados de redes sociais e desenvolvido esforços para a coleta de dados de voluntários que sirvam como regra ouro para aferir a qualidade dos métodos que estamos desenvolvendo. Finalmente, os métodos aqui propostos para enriquecer se-

manticamente dados sobre movimento com dados ligados têm o potencial de alavancar novos desenvolvimentos em armazéns de dados sobre movimento e mineração de dados, explorando a semântica bem definida em conceitualizações e instâncias que ajudam a descrever o mundo em certas ontologias e coleções de dados ligados. Além disso, o enorme volume, a variedade e a velocidade de atualização de coleções de dados sobre movimento (incluindo dados de mídias sociais) e dados ligados, exige técnicas apropriadas para viabilizar tanto métodos como os que estamos desenvolvendo para o enriquecimento semântico de dados de movimento, quanto para o uso dos dados enriquecidos resultantes em análises de informação sobre armazéns de dados e em mineração de dados, assim como a sua utilização em aplicações com tempos de resposta aceitáveis. Assim, buscamos parcerias para utilizar técnicas de processamento paralelo e *big data*.

REFERENCIAS

Alvares, L. O., Loy, A. M., Renso, C., and Bogorny, V. (2011). An algorithm to identify avoidance behavior in moving object trajectories. J. Braz. Comp. Soc., 17(3):193–203.

Arboleda, F. J. M., Fileto, R., and Isaza, F. A. (2010). Season queries on a temporal multidimensional model for olap. Mathematical and Computer Modelling, 52(7-8):1103–1109.

Baglioni, M., de Macêdo, J. A. F., Renso, C., Trasarti, R., and Wachowicz, M. (2009). Towards semantic interpretation of movement behavior. In Sester, M., Bernard, L., and Paelke, V., editors, AGILE Conf., Lecture Notes in Geoinformation and Cartography, pages 271–288. Springer.

Baglioni, M., de Macêdo, J. A. F., Renso, C., and Wachowicz, M. (2008). An ontology -based approach for the semantic modeling and reasoning on trajectories. In Song, I.-Y., Piattini, M., Chen, Y.-P. P., Hartmann, S., Grandi, F., Trujillo, J., Opdahl, A. L., Ferri, F., Grifoni, P., Caschera, M. C., Rolland, C., Woo, C., Salinesi, C., Zimányi, E., Claramunt, C., Frasincar, F., Houben, G.-J., and Thiran, P., editors, ER Workshops, volume 5232 of LNCS, pages 344–353. Springer.

Barbosa, I., Casanova, M. A., Renso, C., and de Macêdo, J. A. F. (2013). Average speed estimation for road networks based on gps raw trajectories. In Hammoudi, S., Maciaszek, L. A., Cordeiro, J., and Dietz, J. L. G., editors, ICEIS (1), pages 490–497. SciTePress.

Battle, R. and Kolas, D. (2012). Enabling the geospatial Semantic Web with Parliament and GeoSPARQL. Semantic Web, 3(4):355–370.

Bogorny, V., Avancini, H., de Paula, B. C., Kuplich, C. R., and Alvares, L. O. (2011). Weka-STPM: a Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization. T. GIS, 15(2):227–248.

Bogorny, V., Renso, C., de Aquino, A. R., de Lucca Siqueira, F., and Alvares, L. O. (2014). CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. T. GIS, 18(1):66–88.

Bogorny, V. and Vinhas, L., editors (2010). XI Brazilian Symposium on Geoinformatics, November 28 to December 01, 2010, Campos do Jordão, São Paulo, Brazil. MCT/INPE.

Bouros, P., Ge, S., and Mamoulis, N. (2012). Spatio-textual similarity joins. Proc. VLDB Endow., 6(1):1–12.

Brilhante, I. R., Berlingerio, M., Trasarti, R., Renso, C., de Macêdo, J. A. F., and Casanova, M. A. (2012). Cometogether: Discovering communities of places in mobility data. In Aberer, K., Joshi, A., Mukherjea, S., Chakraborty, D., Lu, H., Venkatasubramanian, N., and Kanhere, S., editors, MDM, pages 268–273. IEEE Computer Society.

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., and Trani, S. (2013). Learning relatedness measures for entity linking. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R., editors, CIKM, pages 139–148. ACM.

Chen, L., Cong, G., Jensen, C. S., and Wu, D. (2013). Spatial keyword query processing: An experimental evaluation. Proc. VLDB Endow., 6(3):217–228.

Cohen, W. W., Ravikumar, P. D., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In Kambhampati, S. and Knoblock, C. A., editors, IIWeb, pages 73–78.

Cong, G., Lu, H., Ooi, B. C., Zhang, D., and Zhang, M. (2012). Efficient spatial keyword search in trajectory databases. CoRR, abs/1205.2880.

da Rocha, T. R., Willrich, R., Fileto, R., and Tazi, S. (2009). Supporting collaborative learning activities with a digital library and annotations. In Tatnall, A. and Jones, A., editors, WCCE, volume 302 of IFIP Advances in Information and Communication Technology, pages 349–358. Springer.

D'Agostini, C. S. and Fileto, R. (2009). Capturing users' preferences and intentions in a semantic search system. In SEKE, pages 587–591. Knowledge Systems Institute Graduate School.

D'Agostini, C. S., Fileto, R., Dantas, M. A. R., and Gauthier, F. A. O. (2008). Contextual semantic search - capturing, using the user's context to direct semantic search. In Cordeiro, J. and Filipe, J., editors, ICEIS (4), pages 154–159.

de Aquino, A. R., Alvares, L. O., Renso, C., and Bogorny, V. (2013). Towards semantic trajectory outlier detection. In GeoInfo, pages 115–126. MCT/INPE.

De Felipe, I., Hristidis, V., and Rishe, N. (2008). Keyword search on spatial databases. In IEEE 24th Intl. Conf. on Data Engineering, ICDE '08, pages 656–665, Washington, DC, USA. IEEE Computer Society.

Deggau, R., Fileto, R., Pereira, D., and Merino, E. (2010). Interacting with spatial data warehouses through semantic descriptions. In [Bogorny and Vinhas, 2010], pages 122–133.

Fileto, R., Krüger, M., Pelekis, N., Theodoridis, Y., and Renso, C. (2013). Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data. In Ng, W., Storey, V. C., and Trujillo, J., editors, ER, volume 8217 of LNCS, pages 342–355. Springer.

Fileto, R., Medeiros, C. B., Pu, C., Liu, L., and Assad, E. D. (2005). Building a semantic web system for scientific applications: An engineering approach. In Ngu, A. H. H., Kitsuregawa, M., Neuhold, E. J., Chung, J.-Y., and Sheng, Q. Z., editors, WISE, volume 3806 of Lecture Notes in Computer Science, pages 633–642. Springer.

Filho, S. I. V., Fileto, R., Furtado, A. S., and Guembarovski, R. H. (2010). Towards Intelligent Analysis of Complex Networks in Spatial Data Warehouses. In [Bogorny and Vinhas, 2010], pages 134–145.

Furletti, B., Gabrielli, L., Renso, C., and Rinzivillo, S. (2013). Analysis of GSM calls data for understanding user mobility behavior. In Hu, X., Lin, T. Y., Raghavan, V., Wah, B. W., Baeza-Yates, R. A., Fox, G., Shahabi, C., Smith, M., Yang, Q., Ghani, R., Fan, W., Lempel, R., and Nambiar, R., editors, BigData Conference, pages 550–555. IEEE.

Furtado, A. S., Fileto, R., and Renso, C. (2013). Assessing the attractiveness of places with movement data. JIDM, 4(2):124–133.

Gabrielli, L., Rinzivillo, S., Ronzano, F., and Villatoro, D. (2013). From tweets to semantic trajectories: Mining anomalous urban mobility patterns. In Nin, J. and Villatoro, D., editors, CitiSens, volume 8313 of Lecture Notes in Computer Science, pages 26–35. Springer.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti,

R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J., 20(5):695–719.

Leonardi, L., Orlando, S., Raffaetà, A., Roncato, A., Silvestri, C., Andrienko, G. L., and Andrienko, N. V. (2014). A general framework for trajectory data warehousing and visual OLAP. GeoInformatica, 18(2):273–312.

Liu, S., Li, G., and Feng, J. (2012). Star-join: Spatio-textual similarity join. In 21st ACM Intl. Conf. on Information and Knowledge Management, CIKM '12, pages 2194–2198, New York, NY, USA. ACM.

Manco, G., Baglioni, M., Giannotti, F., Kuijpers, B., Raffaetà, A., and Renso, C. (2008). Querying and reasoning for spatiotemporal data mining. In Giannotti, F. and Pedreschi, D., editors, Mobility, Data Mining and Privacy, pages 335–374. Springer.

May, C. and Fileto, R. (2014). Connecting Textually Annotated Movement Data with Linked Data. In IX Regional School on Databases, ERBD, São Francisco do Sul, SC, Brazil (in Portuguese). SBC.

Moreau, E., Yvon, F., and Cappe´, O. (2008). Robust Similarity Measures for Named Entities Matching. In 22nd Intl. Conf. on Computational Linguistics - Volume 1, COLING, pages 593–600, Stroudsburg, PA, USA. Association for Computational Linguistics.

Moreno, F., Pineda, A., Fileto, R., and Bogorny, V. (2014). SMoT+: Extending the SMoT Algorithm for Discovering Stops in Nested Sites. Computing and Informatics, 33(2):327–342.

Nabo, R. G. B., Fileto, R., Renso, C., and Nanni, M. (2014). Annotating Trajectories by Fusing them with Social Media Users' Posts. In Brazilian Symposium on Geoinformatics, GeoInfo, Campos do Jordão, SP, Brazil (submitted).

Nanni, M., Trasarti, R., Renso, C., Giannotti, F., and Pedreschi, D. (2010). Advanced knowledge discovery on movement data with the geopkdd system. In Manolescu, I., Spaccapietra, S., Teubner, J., Kitsuregawa, M., Léger, A., Naumann, F., Ailamaki, A., and Özcan, F., editors, EDBT, volume 426 of ACM International Conference Proceeding Series, pages 693–696. ACM.

Navarro, G. (2001). A guided tour to approximate string matching. ACM Comput. Surv., 33(1):31–88.

Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. In Wainwright, R. L. and Haddad, H., editors, SAC, pages 863–868. ACM.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G. L., Andrienko, N. V., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., de Macêdo, J. A. F., Pelekis, N., Theodoridis,

Y., and Yan, Z. (2013). Semantic trajectories modeling and analysis. ACM Comput. Surv., 45(4). Article 42.

Pelekis, N., Frentzos, E., Giatrakos, N., and Theodoridis, Y. (2008). Hermes: aggregative lbs via a trajectory db engine. In Wang, J. T.-L., editor, SIGMOD Conference, pages 1255–1258. ACM.

Pelekis, N. and Theodoridis, Y. (2014). Mobility Data Management and Exploration. Springer.

Pelekis, N., Theodoridis, Y., and Janssens, D. (2013). On the management and analysis of our lifesteps. SIGKDD Explorations, 15(1):23–32.

Renso, C., Baglioni, M., de Macêdo, J. A. F., Trasarti, R., and Wachowicz, M. (2013). How you move reveals who you are: understanding human behavior by analyzing trajectory data. Knowl. Inf. Syst., 37(2):331–362.

Rinzivillo, S., de Lucca Siqueira, F., Gabrielli, L., Renso, C., and Bogorny, V. (2013). Where Have You Been Today? Annotating Trajectories with DayTag. In SSTD, volume 8098 of LNCS, pages 467–471. Springer.

Rocha, J. A. M. R., Times, V. C., Oliveira, G., Alvares, L. O., and Bogorny, V. (2010). DB-SMoT: A direction-based spatio-temporal clustering method. In IEEE Conf. of Intelligent Systems, pages 114–119. IEEE.

Shen, W., Wang, J., Luo, P., and Wang, M. (2013). Linking named entities in tweets with knowledge base via user interest modeling. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 68–76, New York, NY, USA. ACM.

Simões, D. V., Viana, H., Markey, N., and de Macedo, J. A. F. (2012). Querying trajectories through model checking based on timed automata. In XXV Brazilian Symp. on Databases, SBBD, pages 33–40, São Paulo, SP, Brazil. SBC.

Spaccapietra, S. and Parent, C. (2011). Adding meaning to your steps (keynote paper). In Jeusfeld, M. A., Delcambre, L. M. L., and Ling, T. W., editors, ER, volume 6998 of LNCS, pages 13–31. Springer.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macêdo, J. A. F., Porto, F., and Vangenot,

C. (2008). A conceptual view on trajectories. Data Knowl. Eng., 65(1):126–146.

Wachowicz, M., Ong, R., Renso, C., and Nanni, M. (2011). Finding moving flock patterns among pedestrians through collective coherence. International Journal of Geographical Information Science, 25(11):1849–1864.

Wagner, R., de Macêdo, J. A. F., Raffaetà, A., Renso, C., Roncato, A., and Trasarti, R. (2013). Mob-Warehouse: A semantic approach for mobility analysis with a Trajectory Data Warehouse. In SecoGIS, joint to ER 2013, Hong Kong.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., and Aberer, K. (2013). Semantic trajectories: Mobility data computation and annotation. ACM TIST, 4(3).

Zhang, C., Han, J., Shou, L., Lu, J., and Porta, T. F. L. (2014). Splitter: Mining fine-grained sequential patterns in semantic trajectories. PVLDB, 7(9):769–780.

Zhang, D., Tan, K.-L., and Tung, A. K. H. (2013). Scalable top-k spatial keyword search. In Proceedings of the 16th International Conference on Extending Database Technology, EDBT '13, pages 359–370, New York, NY, USA. ACM.

CONCEPÇÃO DE SISTEMAS INTEGRADOS TOLERANTES A EFEITOS DE RADIAÇÃO

Ricardo Reis e Fernanda Lima Kastensmidt

Abstract.

The sizing scaling of the components of an integrated circuits has increased the their sensibility radiation effects. In the past, only the circuits used in spaceships and satellites were designed including techniques to cope with faults due to radiation effects. Nowadays, the integrated systems used at Earth level are sensible to radiation effects. So, the integrated systems used in critical applications like in medicine, needs the research and developing of techniques to design systems tolerant to the radiation effects. In this paper, it is described some problems and possible solutions to mitigate these problems provoked by transients due to radiation.

Resumo.

A redução das dimensões dos componentes dos circuitos integrados tem provocado o aumento da sensibilidade dos mesmos aos efeitos de radiação. No passado, apenas os circuitos usados em naves espaciais e satélites incluíam técnicas de tolerância a falhas devido a efeitos de radiação. Atualmente, mesmo no nível da Terra, os circuitos integrados são sensíveis aos efeitos de radiação. Com isto, os sistemas integrados usados em aplicações críticas, como aplicações na medicina demandam a pesquisa e desenvolvimento de técnicas que tornem estes sistemas tolerantes aos efeitos de radiação. Neste artigo são descritos alguns dos problemas e soluções possíveis para mitigar os problemas podem ser gerados por transientes gerados devido à incidência de radiação.

1. INTRODUÇÃO

O projeto de sistemas integrados tolerantes a efeitos de radiação era algo considerado no passado apenas para sistemas que eram lançados ao espaço, como naves espaciais e satélites. Atualmente, em função das baixas tensões de alimentação, este problema também afeta sistemas integrados no nível terrestre. Diversos fenômenos ganham relevância à medida em que as dimensões e as tensões de alimentação dos dispositivos semicondutores MOS são reduzidas, fazendo com que o comportamento elétrico destes seja bastante diferente do comportamento de seus semelhantes de geometrias maiores.

Os sistemas de comunicação de satélites, o controle de aeronaves no espaço, a instrumentação biomédica, os servidores de dados, os sistemas de rede e outras aplicações

PGMicro/PPGC, Instituto de Informática – Univ. Federal do Rio Grande do Sul - Porto Alegre - Brasil
reis@inf.ufrgs.br, fglima@inf.ufrgs.br

críticas encontradas em nossa sociedade compartilham de uma característica comum: a necessidade de ser garantida a sua confiabilidade operacional, assim como sua disponibilidade e utilidade, expressão conhecida como *Reliability, Accessability and Serviceability (RAS)*. A confiabilidade é a habilidade de um sistema operar continuamente sem erros e de manter a integridade de dados mesmo na presença de defeitos e falhas. A disponibilidade do sistema é a porcentagem do tempo que um sistema permanece acessível para uso. A utilidade relaciona-se ao tempo necessário para restaurar o sistema ao serviço devido a ocorrência de uma falha. Porém, garantir a capacidade de RAS de um sistema computacional torna-se cada vez mais desafiador nas tecnologias nanométricas. Isto deve-se ao fato dos sistemas computacionais crescerem em complexidade, possuindo cada vez mais transistores e funcionalidades, o que impossibilita a realização de testes exaustivos. Além disso, os dispositivos que compõem os sistemas computacionais são cada vez mais sensíveis a flutuações de comportamento elétrico. Isto faz com que, hoje, mesmo os sistemas que operam próximo ao nível do mar devam ser projetados aplicando técnicas de tolerância a efeitos de transientes e de tolerância à variabilidade dos processos de fabricação.

Além disso, há o efeito de "*aging*", ou seja, o envelhecimento do circuito, que torna-se mais eminente nas tecnologias nanométricas [Vasquez 2012]. Um dos efeitos mais importantes é conhecido como NBTI ("*Negative Bias Temperature Instability*") que altera a tensão de limiar (*threshold*) dos transistores PMOS, degradando o funcionamento do transistor. A fim de compreender melhor as características destes efeitos (variabilidade e "*aging*") em tecnologias nanométricas, é necessário desenvolver meios exatos e eficientes de medir e caracterizar seus efeitos. Atualmente, isto limita em muito a validade dos modelos e metodologias empregados no projeto de circuitos integrados, dificultando avanços nas tecnologias de processamento físico-químico do silício que garantam a qualidade e lucratividade dos produtos.

As flutuações no comportamento elétrico de um circuito podem ter também origem externa. Os circuitos no estado-da-arte tornam-se a cada dia mais vulneráveis a ruídos externos e internos, ao acoplamento de sinal e à radiação presente não apenas no espaço, mas também na atmosfera terrestre [Velazco, Fouillat, Reis 2007]. Essa vulnerabilidade deve-se à evolução das tecnologias de fabricação que tem resultado em uma contínua diminuição nas dimensões dos transistores e na tensão de alimentação, em aumento de velocidade de processamento e de densidade lógica, e em diminuição da carga crítica necessária para a ocorrência de falhas no circuito. As falhas geradas por partículas energizadas são conhecidas como falhas de efeito transiente ("*soft errors*"). As falhas transientes geram pulsos transitórios de tensão que podem acontecer em nós do circuito e que podem ser armazenados em elementos de memória, aparecer na saída do circuito ou afetar diretamente um *latch* ou *flip-flop* interno gerando um *bit-flip*. As falhas transientes, antes relevantes apenas para aplicações aeroespaciais, são hoje

importantes também para sistemas que operam ao nível do mar, devido à contínua diminuição da dimensão e tensão de alimentação dos dispositivos que torna os circuitos sensíveis a transientes com menores cargas. Além disso, a presença do circuito em um fluxo de cargas energizadas pode ocasionar, a longo prazo, degradação elétrica e falha permanente devido à dose total de ionização ("*Total Ionization Dose*" ou TID).

Nas tecnologias nanométricas atuais e futuras, uma questão fundamental é o desafio de desenvolver circuitos integrados confiáveis com número cada vez maior de dispositivos, os quais são, individualmente, cada vez menos confiáveis. Nestas tecnologias aumenta-se consideravelmente a sensibilidade a flutuações como:

• ruídos internos e externos ao circuito, cujo efeito são falhas transientes ou intermitentes.

• variabilidade do processo de fabricação, cujo efeito é permanente para uma determinada área do circuito.

• efeito de envelhecimento precoce ("*aging*").

• ionização devido a partículas alfa geradas por nêutrons presentes na atmosfera terrestre, cujo efeito são falhas transientes na lógica combinacional e sequencial.

Em suma, faz-se necessário o projeto de sistemas confiáveis usando componentes não confiáveis. Os sistemas computacionais variam em sua complexidade e portabilidade, mas todos possuem um conjunto de hardware e software trabalhando de maneira integrada para a realização de uma tarefa. O grande desafio é manter o sistema funcionando de forma adequada mesmo na presença de defeitos e flutuações que podem ocorrer durante sua vida útil. Para garantir este requisito é necessário pesquisar métodos, técnicas, modelos, dispositivos e padrões de arquitetura capazes de auxiliar os projetistas e desenvolvedores de grandes sistemas de software e hardware a atingirem esses objetivos.

Projetando e fabricando um circuito integrado mais confiável, é possível posteriormente desenvolver soluções de confiabilidade em software com um menor custo e menor penalidade em desempenho. Assim, pode-se otimizar custos como área e potência, e maximizar desempenho e funcionalidade ao garantir que ambos, software e hardware, tenham a preocupação de se manter confiáveis, disponíveis e úteis (RAS) ao longo de sua vida.

A analise simultânea dos efeitos de variabilidade, *aging*, ruído em baixa frequência, soft erros e TID é inédita. Não há na literatura e nem comercialmente ferramentas e metodologias que propõem analisar esses efeitos concomitantemente. Porém essa analise é fundamental para sistemas computacionais usando circuitos fabricados em tecnologia nanométricas possam manter-se operando de maneira correta ao longo de sua vida útil.

A detecção e diagnóstico das falhas e efeitos de variabilidade e *aging* são fundamentais para guiar os projetistas no desenvolvimento de sistemas confiáveis. A etapa de teste destes sistemas precisa tratar do uso de "*Design for Testability*" (DFT) e "*Design for Manufacturability*" (DFM) para cedo identificar falhas e poder tolerá-las. As técnicas para tolerar variabilidade são baseadas no dimensionamento de transistores, em árvores de relógio, posicionamento e roteamento, etapas que podem ser integradas em ferramentas de CAD. Técnicas para detectar o efeito de *aging* são baseadas em sensores embarcados compostos por "*delay-locked loops*" (DLL) e "*phase-locked loop*" (PLL), que podem também funcionar como atuadores na tensão de alimentação para diminuir o efeito do *aging* [Vasquez 2012]. As técnicas de tolerância a falhas transientes podem ser implementadas na descrição arquitetural e lógica do circuito, assim como no nível elétrico e de leiaute. As técnicas normalmente são baseadas em redundância espacial ou temporal. O desafio é desenvolver técnicas que tolerem falhas transientes em múltiplos nós do circuito e de longa duração.

Queremos também observar que em 2006 a SBC publicou um texto indicando os 5 grandes desafios para a pesquisa avançada em Computação até 2016 [SBC 2006]. Este artigo está relacionado ao desafio "Impactos para a área de computação da transição do silício para novas tecnologias" mas considerando o dito na publicação que fizemos em [Bampi 2009] de que esta transição não deverá ocorrer nas próximas décadas e que a tecnologia CMOS ainda possui ao menos muitas décadas de vida. Esta afirmação também está amparada no ITRS (International Technology Roadmap for Semiconductors) [ITRS2009], que claramente mostra que a tecnologia CMOS ainda será a tecnologia principal para a implementação de sistemas integrados durante várias décadas. Cabe observar que este *roadmap* é elaborada por especialistas de empresas, centros de pesquisa e universidades que estão na ponta dos avanços tecnológicos do setor de semicondutores. Estes *roadmaps* vem sendo desenvolvidos desde 1992, tendo sido demonstrado ao longo do tempo uma significativa precisão, mas os avanços da tecnologia CMOS tem sido mais significativos do que os previstos nos primeiros *roadmaps*.

2. APLICAÇÕES NA SAÚDE E EM OUTRAS ÁREAS.

Cabe observar que as aplicações de sistemas integrados confiáveis e tolerantes à efeitos de radiação aplica-se não apenas às áreas aeroespaciais e aeronáuticas, mas também à área automotiva, à medicina (especialmente sistemas críticos implantados ou que monitoram parâmetros de vida), e diversas outras áreas que necessitam sistemas com altíssima confiabilidade.

Na área médica, mais e mais aplicações utilizam sistemas eletrônicos e computacionais que devem ter o máximo grau de confiabilidade. Podemos classificar em sistemas

que são utilizados por um paciente e sistemas que são usados para efetuar intervenções em um paciente ou para monitoramento externo. No caso de sistemas utilizados por um paciente, é desejado uma miniaturização dos mesmos. Com, isto o sistema pode operar com uma demanda menor de energia, mas também com um maior grau de confiabilidade. A integração de um sistema em apenas um chip conduz a uma solução com maior grau de confiabilidade. A miniaturização e a redução da energia para o funcionamento de circuitos integrados tem aumentado o desenvolvimento de sistemas integrados que possam ser implantados em seres humanos. Estes sistemas, devem naturalmente ter um máximo grau de confiabilidade, pois qualquer falha pode ter consequências de alta gravidade. Portanto, devem incorporar técnicas para serem tolerantes aos efeitos de radiação.

Equipamentos de monitoração médica ou usados em intervenções cirúrgicas também devem incorporar técnicas de tolerância a falhas, inclusive às devido a efeitos de radiação. Portanto, é evidente que é estratégico para o Brasil desenvolver aplicações para a medicina que sejam tolerantes à falhas.

O Brasil encontra-se em uma região do globo terrestre com maior incidência de partículas na atmosfera devido a Anomalia do Atlântico Sul (SAA). Logo, como pode ser visto na Figura 1, a taxa de erros em satélites por exemplo, do inglês *soft error rate* (SER) é muito maior e mais frequente na região da anomalia, devido ao formato e intensidade do campo magnético da Terra nesta região.

Por isso, faz-se necessário ter projetos que tenham como objetivo investigar em detalhes essa anomalia e seus efeitos em aeronaves comerciais, militares e satélites por exemplo. Com esse objetivo, surgiu o projeto NanoSat-BR1 (Figura 2). O projeto desse nano satélite foi coordenado pelo INPE e teve como universidades parceiras ao projeto as Universidade Federal de Santa Maria e a Universidade Federal do Rio Grande do Sul.

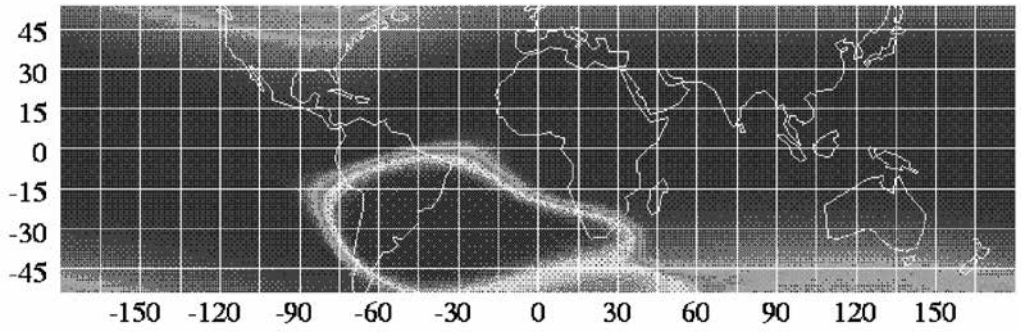


Figura 1. Soft Error Rate at Earth (Nasa 2002)

A UFRGS foi responsável pelo desenvolvimento de parte da carga útil com a implementação de um processador embarcado MIPS de 32-bits em um circuito programável FPGA programado por memoria flash (FPGA Actel ProASIC3E da empresa Microsemi) e de um software tolerante a falhas capaz de detectar falhas transientes neste processador.

Figura 2. NanoSat-BR1 incluindo chips desenvolvidos pela UFRGS e UFSM

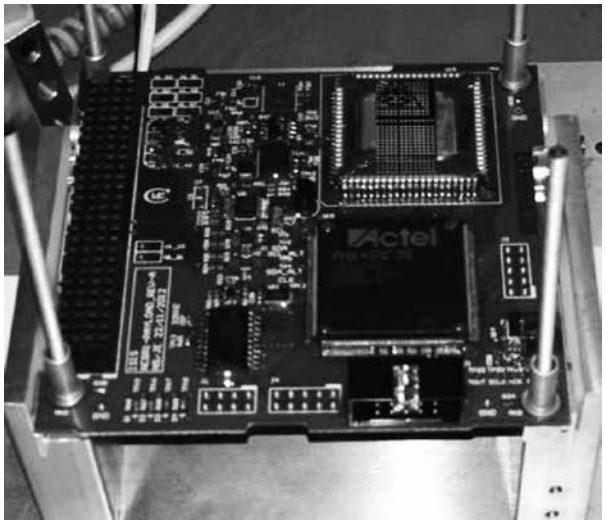


Figura 3. Detalhe da placa de carga útil

3. FALHAS TRANSIENTES EM TECNOLOGIAS NANOMÉTRICAS E DESAFIOS

Falhas induzidas por radiação vêm se tornando uma forte ameaça ao correto funcionamento dos dispositivos eletrônicos comerciais não somente operando no espaço mas também na Terra devido a interação dos neutrons com o material. A compreensão dos efeitos da radiação nos dispositivos eletrônicos tornou se importante, especialmente no caso de aplicações espaciais, aviônicas e militares, já que a exposição destes circuitos à partículas energéticas e/ou fótons pode resultar em efeitos graves na aplicação. Diversos casos de mau funcionamento ou falhas terminais de aplicações aeroespaciais já foram relatados (Velazco, Fouillat and Reis 2007), mais recentemente o caso ocorrido em 2008 onde uma aeronave Airbus 330 efetuou duas descidas abruptas sucessivamente, uma de 150m e outro de 120m, ferindo seriamente 12 pessoas, a causa do incidente foi uma falha no sistema de computador de bordo, acredita-se que induzida por raios cósmicos.

Existem dois tipos de interação entre a partícula e o material semicondutor: a ionização direta, gerada pela própria partícula, e a ionizante indireta, gerada por partículas secundarias derivadas da reação entre a partícula primaria e o material colidido. Essa ionização, direta ou indireta, gera um acumulo de carga no que é coletado pelo nó atingido, gerando uma perturbação no nível de tensão do mesmo.

O primeiro efeito transiente é chamado de *Single Event Effects* (SEE) e são causados quando partículas energéticas presentes no espaço, como prótons, elétrons e íons pesados (prótons), colidem com uma área sensível do circuito eletrônico, depositando carga na região da junção p-n do transistor. Ou quando nêutrons presentes na atmosfera terrestre colidem com o material semicondutor provocando partículas secundarias (normalmente do tipo alfa) e essas ionizam o material, depositando certa carga na junção p-n. Dependendo de uma série de fatores um SEE pode gerar um efeito não observável, perturbar a operação do circuito de forma transiente, mudar um estado lógico ou até mesmo causar danos permanentes ao dispositivo eletrônico (Dodd and Massengil 2003). Os SEE não destrutivos, são chamados de *single event upsets* (SEU) e *single event transients* (SET).

Existem duas formas pelas quais uma partícula energética pode depositar cargas em um dispositivo semicondutor: ionização direta pela partícula em si, Figura 4(a), e ionização por partículas secundarias geradas pela colisão entre a partícula incidente e o material atingido, Figura 4(b). Os dois mecanismos podem levar a erros no funcionamento de um circuito devida à carga coletada pelo material atingido.

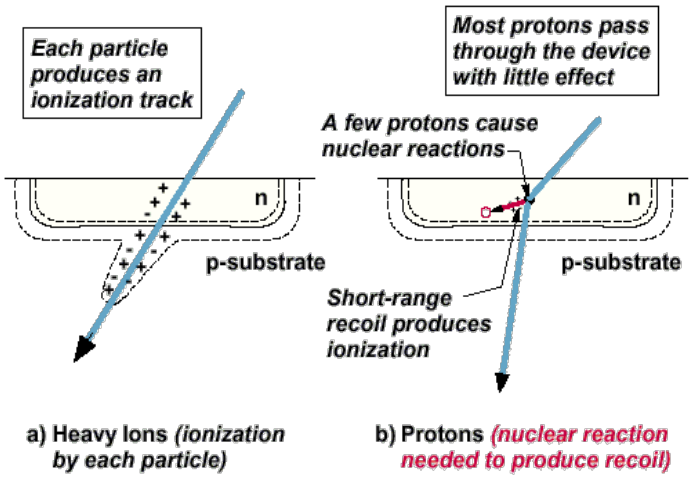


Figura 4. Mecanismos de deposição de carga de um SEE: (a) ionização direta; (b) ionização indireta.

Quando uma partícula energética colide com um material semicondutor, ela liberta pares de elétron-lacuna pelo caminho conforme perde sua energia. A partícula energética repousa no material ao esgotar toda sua energia, a geração de pares elétron-lacuna e a posterior coleta de carga em decorrência do evento, resulta em um pulso de corrente no dispositivo. O formato desse pulso decorre de dois mecanismos distintos que ocorrem após a passagem da partícula incidente: inicialmente por deriva (*drift*) e, posteriormente, por difusão (*diffusion*). Frequentemente se usa transferência linear de energia (LET, *linear energy transfer*) para descrever o deposito de energia por unidade de distância de uma partícula quando a mesma atravessa um material.

Apesar da ionização direta por partículas leves normalmente não depositar carga suficiente para causar SEE isto não significa que essas partículas podem ser ignoradas. Tanto prótons quanto nêutrons podem produzir falhas transientes através da ionização indireta. Conforme nêutrons e prótons, altamente energizados, atravessam o semicondutor eles podem colidir com os núcleos do material gerando algumas reações nucleares.

O produto de qualquer uma destas reações nucleares pode depositar carga por onde atravessam, por meio da ionização direta. Essas partículas geradas são muito mais pesadas que o próton ou o nêutron que as gerou, desta forma elas depositam uma densidade de carga maior enquanto viajam sendo assim capazes de gerar SEE.

Um fenômeno observado nas tecnologias nanometricas é o *Single-event charge sharing*, onde a deposição de carga ocorre em diversos transistores do circuito, que podem pertencer a diversos portas lógicas do circuito, gerando falhas múltiplas, como ilustra a Figura 5.

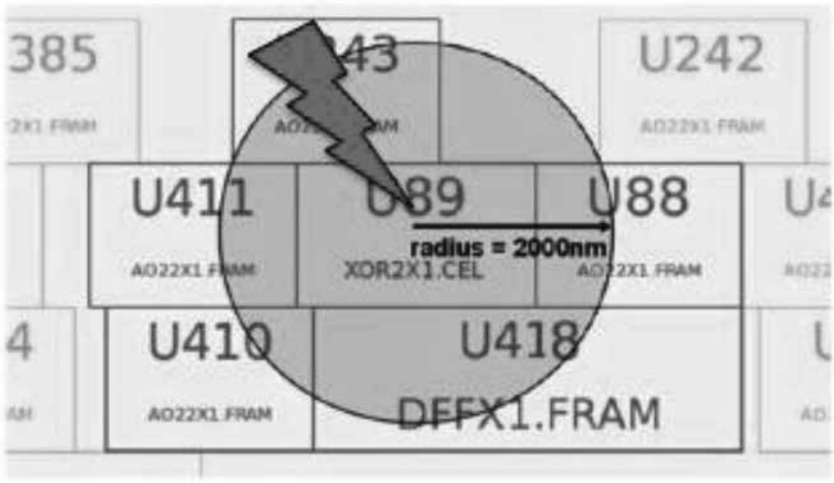


Figura 5. *single-event* induzindo coleta de carga múltipla nas células adjacentes no leiaute

Normalmente as junções p-n com polaridade reversa são as áreas mais sensíveis em um circuito integrado quando se diz respeito a falhas causadas por partículas energéticas, principalmente se a junção está flutuando ou sendo carregada de forma fraca (quando o transistor que fornece a corrente, e por consequência a tensão do nó, é pequeno, tornando mais difícil a manutenção do estado do nó). O forte campo presente na zona de depleção da junção p-n, polarizada de forma reversa, pode coletar de forma muito eficiente as cargas induzidas por partículas através do processo de deriva (*drift*) gerando a uma corrente transiente no local.

A **Figura** ilustra os processos que ocorrem durante a colisão de uma partícula energética com uma junção p-n. Quando o caminho ionizado resultante atravessa ou passa perto da região de depleção os portadores são coletados rapidamente pelo campo elétrico gerando uma perturbação transiente de corrente/tensão no nó (**Figura** a). Uma característica importante é a distorção que ocorre no potencial eletrostático da região de depleção na forma de funil. Esse funil aumenta significativamente a coleta de carga por deriva ao aprofundar o campo elétrico no substrato (**Figura** b). Uma carga extra é coletada conforme os elétrons difundem-se na região de depleção até todos os portadores serem coletados, recombinados ou difundidos (**Figura** c). O gráfico na **Figura** d

elucida a diferença na coleta de carga por deriva e por difusão, no caso da coleta por deriva a mesma ocorrem dentro de nanosegundo, já a difusão ocorre numa escala de tempo maior (centenas de nanosegundos).

Single event upset (SEU), evento de inversão único em tradução livre, é definido pela inversão do valor armazenado no elemento de memória, essa inversão de valores é comumente chamada de *bit-flip*. Esta falha é temporária, já que pode ser corrigida na próxima captura do elemento de memória, porém esta falha, se propagada, pode gerar erro na execução do circuito. Um SEU ocorre quando uma partícula colide com uma área sensível do elemento de memória e deposita uma carga mínima suficiente no material para ocasionar a inversão do valor armazenado. Esse elemento pode ser uma memória dinâmica (DRAM) memória estática (SRAM) ou um flip-flop, e a carga mínima para ocasionar a inversão do valor armazenado é chamada de carga critica (Q_{crit}). O SEU é um fenômeno reversível, o estado da célula pode ser recuperado por uma operação de escrita usual. A taxa de falhas de SEU, a SER (*soft error rate*), normalmente é expressa em falhas no tempo (FIT, *failure in time*), a quantidade de falhas em 10^9 horas (um bilhão de horas).

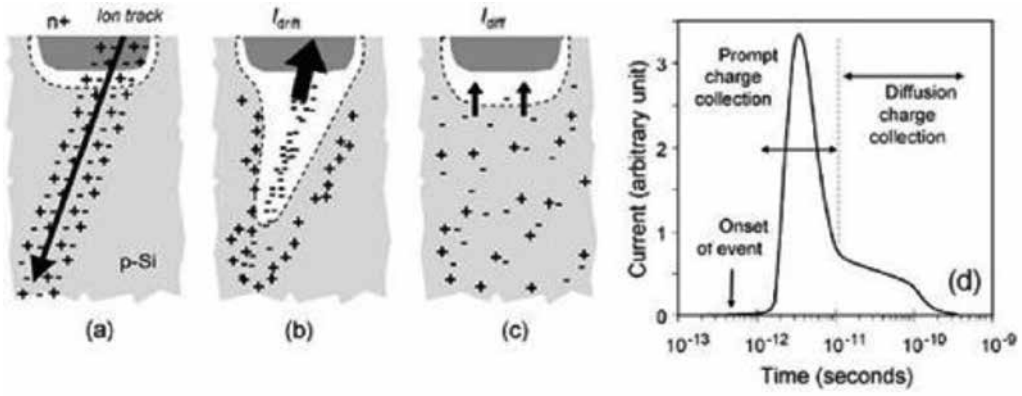


Figura 6. Ilustração da coleta de carga em uma junção p-n de silício (a) imediatamente após a colisão de um íon pesado, (b) durante a coleta de carga (c) durante a coleta de carga por difusão (d) gráfico da corrente induzida em função do tempo (Baumann 2005)

Single Event Transient (SET) (Figura 7), evento transiente único ou falha transiente única, em tradução livre, constitui uma perturbação de tensão/corrente transiente gerada quando uma partícula energética atinge um nó contido numa parte combinacional do circuito integrado. Com a miniaturização constante do tamanho da tecnologia CMOS já ficou claro que o SET tornou-se um mecanismo significativo nas taxas de erro. O escalonamento da tecnologia veio acompanhando de maiores frequências de operação, menores tensões de alimentação e margens de ruídos menores, tornando maior a sensibilidade do circuito a SET.

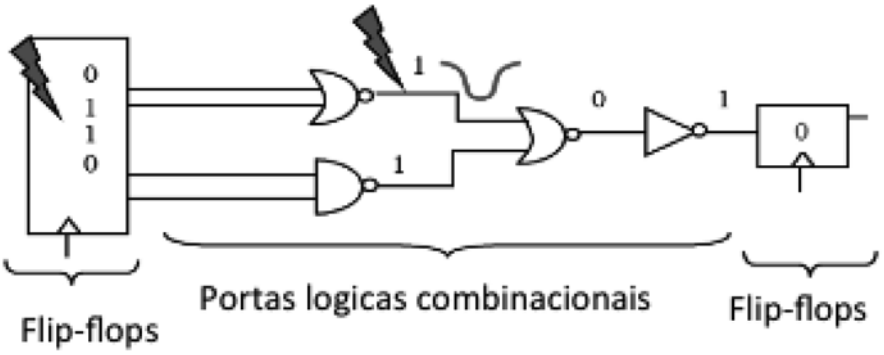
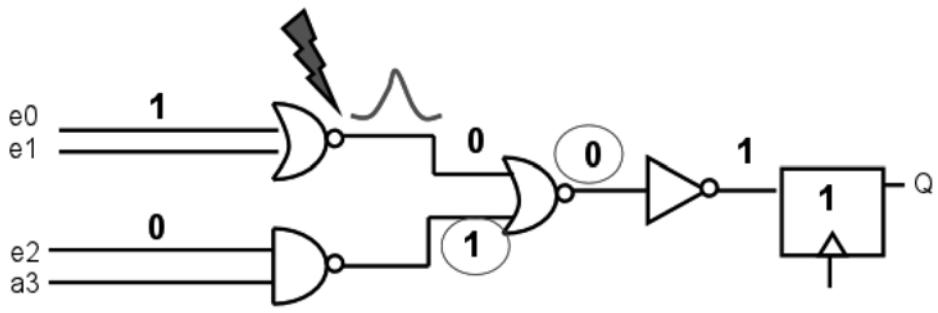


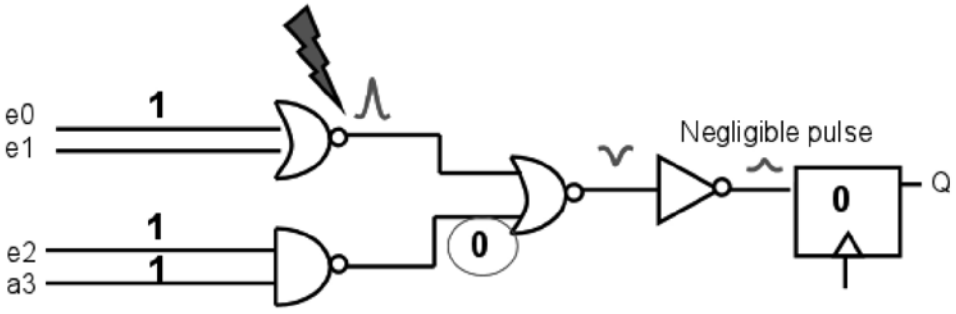
Figura 7. SEU é um bit-flip que aconcete nos flip-flops e o SET é um pulso transiente que acontece no dreno dos transistores desligados das portas logicas combinacionais

Qualquer nó do circuito combinacional pode ser afetado por uma falha e gerar uma perturbação transiente na tensão de amplitude e duração suficiente opera ser propagada ao longo do circuito combinacional até ser capturada por um elemento de memória. No entanto somente alguns transientes são capturados, a chance de um SET ser capturado envolve três questões: a probabilidade de existir um caminho funcionalmente sensível ao SET entre o nó e o elemento sequencial; a taxa com a qual o SET perde força a cada nível lógico que atravessa até alcançar o elemento sequencial; e a chance do pulso transiente gerado ser efetivamente capturado e armazenado no elemento sequencial. Essas três questões levam a três fenômenos de mascaramento:

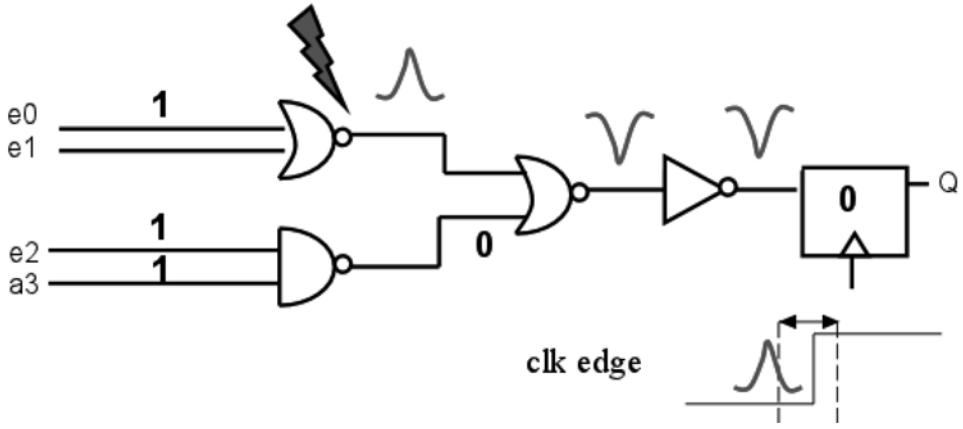
O mascaramento lógico ocorre quando uma falha afeta um nó que não é capaz modificar a saída da porta lógica seguinte. Por exemplo, quando um SET propaga até a entrada da porta lógica NOR (NAND), mas uma das outras entradas é quem controla o estado da saída, no caso de uma NOR (NAND) a entrada esta em 1(0), o SET será completamente mascarado e a saída permanecera intocada. A Figura 8A elucida a questão.



mascaramento logico



mascaramento elétrico



mascaramento temporal ou de captura

Figura 8. Mascaramentos de SET (Kastensmidt 2003)

Esse tipo de mascaramento quando o pulso elétrico gerado pelo SET é atenuado conforme ele se propaga por outras portas lógicas. Esse efeito de atenuação ocorre com os transientes que possuem uma largura de banda maior que a frequência de corte da porta lógica (Munteanu and Autran 2008). Nestes casos a amplitude do transiente pode ser reduzida e seus tempos de subida e descida aumentados e, eventualmente, o pulso pode desaparecer. No entanto, em alguns casos pulsos de baixa frequência com grandes amplitudes podem acabar sendo amplificados. A Figura 8B elucida o efeito de mascaramento elétrico em uma cadeia de inversores.

O mascaramento temporal, ou mascaramento por janela de captura (*latching-window masking*), ocorre quando o SET, resultante da colisão de uma partícula energética, propaga até um circuito sequencial porém não atende os requisitos temporais (bordas

relógio e tempos de *setup* e *hold*) necessários para ser capturado. A Figura 8C elucida um caso de mascaramento temporal no qual o SET chega atrasado em relação a borda de relógio (*CK*), não sendo capturado.

A Figura 9 mostra o efeito da taxa de erro (do ingles, soft error rate – SER) com a diminuição da tensão de alimentação. Note que o SER aumenta com a diminuicao da tensao para todas as tecnologias.

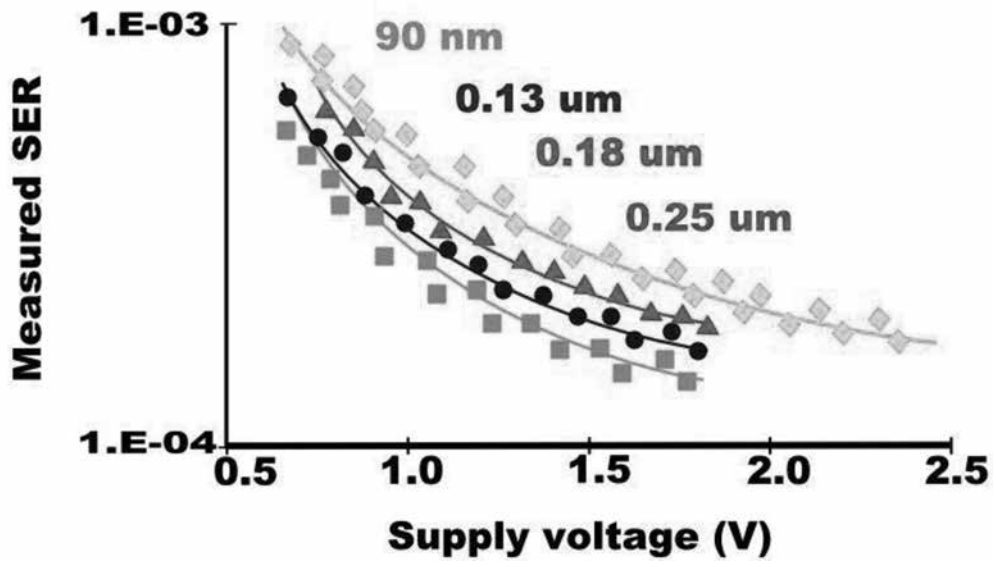


Figura 9. PMOS Diode SER vs. Supply voltage (HAZUCHA, 2003)

4. TÉCNICAS DE MASCARAMENTO E CORREÇÃO DE FALHAS TRANSIENTES

As técnicas de mascaramento de falhas podem ser aplicadas em diversos níveis de abstração do projeto de circuito integrado como mostra a Figura 10. As técnicas de mascaramento são baseadas em redundância logica normalmente e no uso de votador de maioria para selecionar o correto valor na presença de erro. As técnicas de triplicação e votação por exemplo podem ser implementadas no nível de arquitetura RTL por exemplo como mostra a Figura 10 (representação 1). A correção das falhas pode acontecer com a inicialização do sistema, com votadores em laço ou com alguma outra técnica em nível lógico para corrigir a falha. Neste nível, blocos inteiros de logicas podem ser triplicados por exemplo.

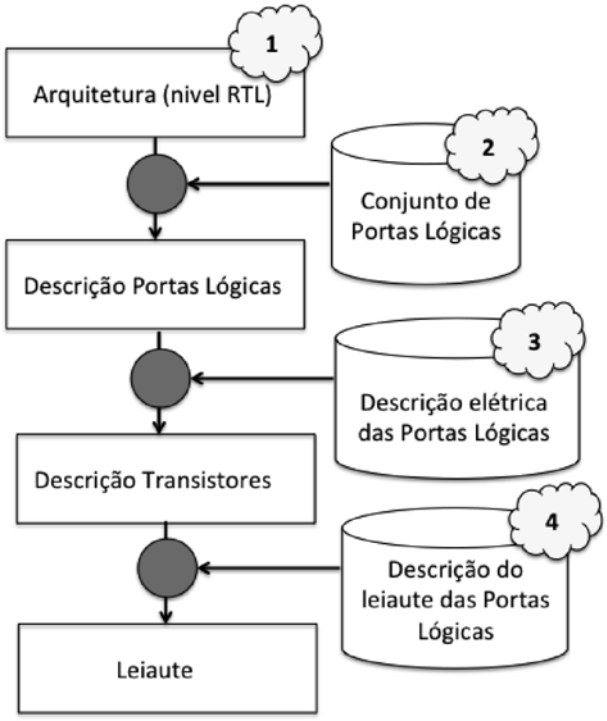


Figura 10. Níveis de abstração simplificados de um projeto de circuito integrado

No nível de porta logica (Figura 10 – representação 2), a redundância é implementada em uma menor granularidade, a granularidade de porta logica. Os flip-flops podem ser substituídos por flip-flips triplicados com votador, as portas logicas combinacionais podem ser substituídas por outras portas logicas compostas por redundância, paridade e votadores (Figura 11).

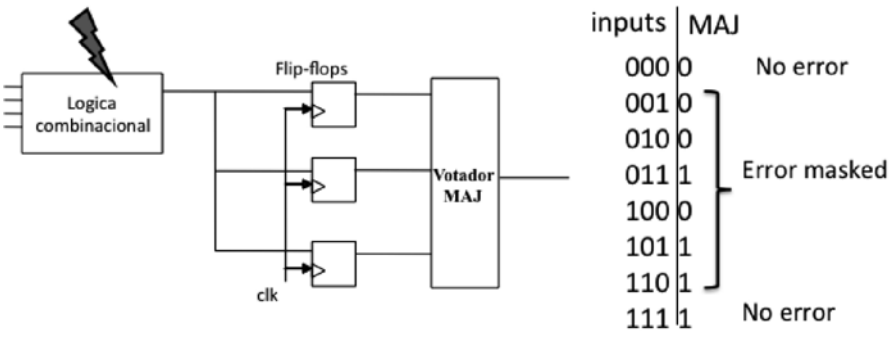


Figura 11. Exemplo de TMR, com triplicação de FF com votador

No nível de transistor, Figura 10 (representação 3), a redundância é implementada em nível de transistor. Por exemplo os flip-flops podem ser substituídos por flip-flop do tipo robusto chamado DICE (Figura 12) (Calin 1996).

No nível de leiaute, Figura 10 (representação 4), técnicas de leiaute são empregadas usando anéis de proteção (Figura13) para redução do compartilhamento de cargas (single-event charge sharing), aumento de transistores, formato dos transistores, uso extra de contatos, e outras técnicas para reduzir a coleta de carga na porta logica sob o efeito da ionização.

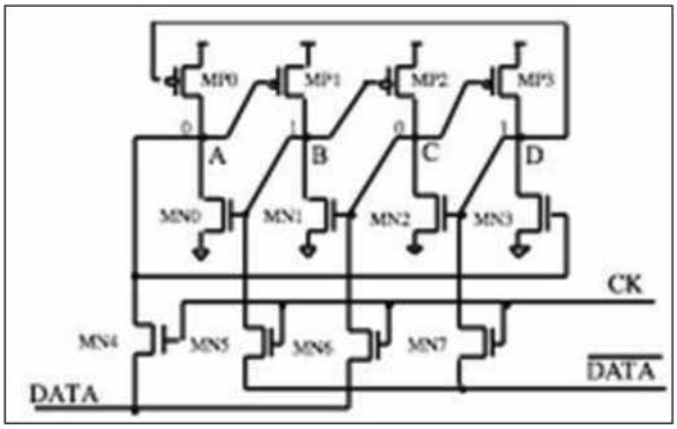


Figura 12. Exemplo de latch robusto a SEU chamado de célula DICE (Calin 1996)

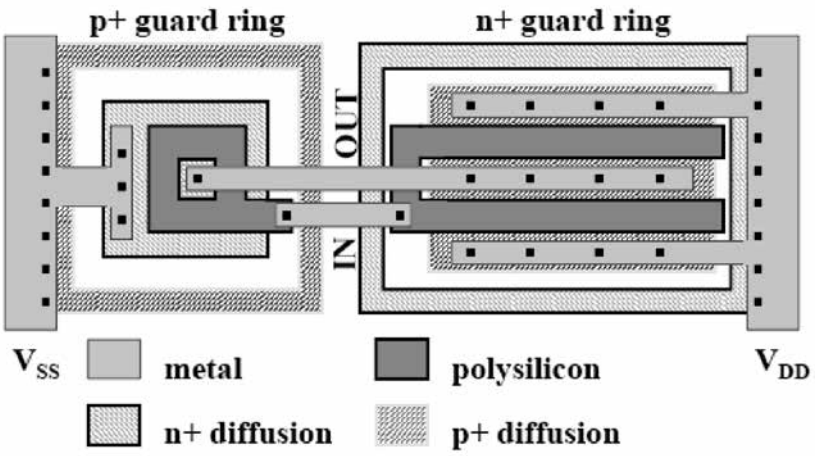


Figura 13. Inversor usando ELT e Anéis de Guarda (ANELLI, 2000)

Todas as técnicas, independente do nível de abstração, que são implementadas apresentam um acréscimo na área, uma diminuição no desempenho do circuito e aumento da potência quando comparado ao circuito original. A escolha de uma técnica ou outra se resume no comprometimento entre o custo extra em área, desempenho, potência e cobertura no mascaramento das falhas.

4. EFEITOS DE VARIABILIDADE E ENVELHECIMENTO NA CONFIABILIDADE

As nanotecnologias CMOS, em duração da redução das dimensões de seus componentes são cada vez mais sensíveis aos efeitos de variabilidade (Neuberger 2009) (Neuberger 2014) (Kastensmidt 2014) (Vasquez 2012). Existem vários tipos e fontes de variabilidade. Um tipo de variabilidade é devido às variações no processo de fabricação, outra fonte de variabilidade é devido ao ambiente (temperatura, tensão de alimentação,..) e também pode ser devido a efeitos de envelhecimento. As fontes de variabilidade devido ao processo de fabricação podem ter origem na etapa de litografia, na dopagem, na espessura de camadas, no polimento físico-químico, e outros. Na litografia as variações podem ser, por exemplo, devido a erros de dose, problemas de foco na ótica, erros na máscara ou erros de superposição de figuras. Em relação a efeitos de envelhecimento, um deles é devido a efeitos de eletromigração provocados especialmente por altas densidades de corrente no funcionamento do chip (Possser 2014). A mudança da posição do pino de saída de uma célula lógica pode representar uma maior ou menor densidade de corrente nas linhas metálicas, o que pode aumentar ou diminuir os efeitos de eletromigração (Figura 14).

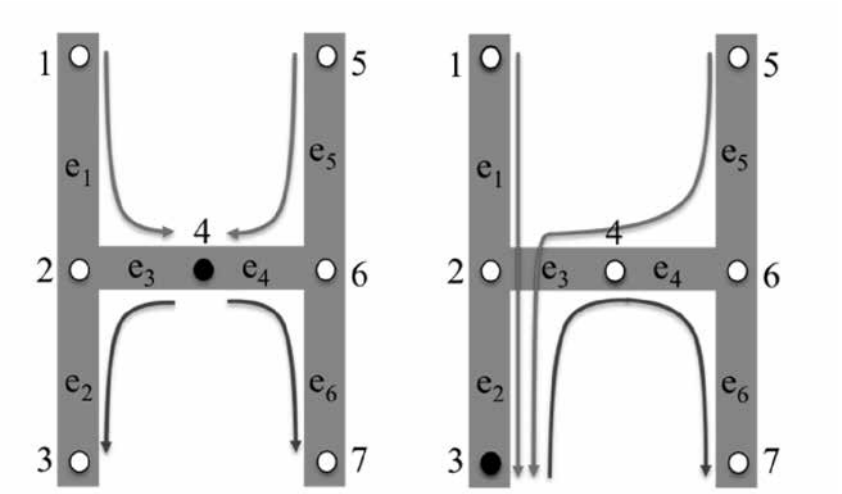


Figura 14. Mudança de densidade de corrente pela mudança de posicionamento do pino de saída (Posser, 2014)

5. CONCLUSÕES

As áreas de saúde e medicina, assim como as áreas aeroespacial, transporte e outras áreas que são de aplicações críticas demandam o projeto de sistemas integrados que tenham o máximo grau de confiabilidade. A integração de sistemas computacionais e eletrônicos em um único chip é um passo importar para aumentar a confiabilidade destes sistemas. Além disto a diminuição das dimensões dos componentes de circuito integrado e das tensões de alimentação tornam os circuitos mais sensíveis a transientes. Ou seja, a amplitude de um transiente necessário para provocar uma falha passa a ser menor. Ou seja, a energia gerada por uma partícula que incida sobre o dreno ou fonte de um transistor, necessária para provocar uma falha é cada vez menor. Portanto, devem ser incorporadas técnicas de tolerância à falhas devido a efeitos de radiação.

Portanto, a geração de circuitos integrados com altas taxas de confiabilidade, e que sejam tolerantes a efeitos de radiação é certamente um dos Grandes Desafios da Computação para as próximas décadas.

6. REFERÊNCIAS

[Anelli 2000] ANELLI, G.M. **Conception et Caracterisation de Circuits Integres Resistants aux Radiations Pour lgheles Detecteurs de Particules du LHC en Technologies CMOS Submicroniques Profondes**. 2000. Thesis (PhD in Laboratoire Européen pour la Recherche Nucléaire - Institut National Polytechnique De Grenoble, France.

[Baumann 2005] BAUMANN, R. **Radiation-Induced Soft Error in Advanced Semiconductor Technologies**. IEEE TRANSACTIONS ON DEVICE AND MATERIALS RELIABILITY. [S.l.:s.n], v.5, n.3, Sep. 2005

[SBC 2006] Brazilian Computer Society. **Grand Challenges in Computer Science Research in Brazil 2006-2016**, 25pgs. In: http://www.sistemas.sbc.org.br. Último acesso em 22 de março de 2010.

[Bampi 2009] BAMPI, S., SUSIN, A., REIS, R., **Systems Architectural Challenges for Transitional and Compatible to CMOS Technologies in Giga-Scale Hardware Integration**, SEMISH 2009, Anais do 36º Seminário Integrado de Software e Hardware, Bento Gonçalves, 21 a 22 de Julho de 2009, p. 281-292, ISSN: 2175-2761.

[Calin 1996] CALIN, T.; NICOLAIDIS, M.; VELAZCO, R., **Upset Hardened Memory Design for Submicron CMOS Technology**. In: IEEE Transactions on Nuclear Science. VOL. 43, NO. 6, December 1996.

[ITRS 2009], International Roadmap Committee. "**The International Technology Roadmap for Semiconductors - 2009**". In http://www.itrs.net/home.html. Último acesso 26 de março de 2010.

[Hazucha, 2003] HAZUCHA, P. et al. Neutron soft error rate measurements in a 90-nm CMOS process and scaling trends in SRAM from 0.25-um to 90-nm generation. IEEE Electron Devices Meeting. [S.l:s.n], 2003.

[Velazco, Fouillat, Reis 2007] VELAZCO, R , FOUILLAT, P, REIS, R., **Radiation Effects on Embedded Systems**, Springer, June 2007. ISBN 978-1-4020-5645-1

[Neuberger 2009] NEUBERGER, G., WIRTH, G., REIS, R., **Protecting Digital Circuits Against Hold Time Violation Due to Process Variability**, In: Chip on The Dunes, 2009, Natal. 31 August - 3 September 2009, 22nd Symposium on Integrated Circuits and System Design, ACM Press, 2009. ISBN: 978-1-6055-8705-9

[Neuberger 2014] NEUBERGER, G., WIRTH, G., REIS, R., **Protecting Chips Against Hold Time Violations Due to Variability**, Springer, 107 p., 2014. ISBN 978-94-007-2426-6. DOI 10.1007/978-94-007-2427-3

[Bastos 2009] BASTOS, R., KASTENSMIDT, F., REIS, R. **Design of a Soft-Error Robust Microprocessor**, Microelectronics Journal, V. 40, N. 7, Elsevier Publishers, ISSN: 0026-2692, July 2009, p. 1062-1068. DOI: 10.1016/j.mejo.2008.10.001

[Kastensmidt 2014] KASTENSMIDT, F., TONFAT, J., BOTH, T., RECH, P., WIRTH, G., REIS, R., BRUGUIER, F., BENOIT, P., TORRES, L., FROST, C., **Voltage scaling and aging effects on soft error rate in SRAM-based FPGAs**, Microelectronics Reliability, Volume 54, Issues 9–10, September–October 2014, Pages 2344-2348, published by Elsevier B.V. in 2014. ISSN: 0026-2714, DOI: 10.1016/j.microrel.2014.07.

[Posser 2014] POSSER, G., MISHRA, V., JAIN, P., REIS, R., SAPATNEKAR, S., A Systematic Approach for Analyzing and Optimizing Cell-Internal Signal Electromigration, ICCAD 2014 – 33rd IEEE/ACM International Conference on Computer-Aided Design, November 3-6, San Jose, USA. p. 486-491, ISBN: 978-1-4799-6278-5

[Reis, Co, Wirth, 2015] REIS, R., CAO, Y., WIRTH, G., Circuit Design for Reliability, Springer, 274 p., 2015, ISBN 978-1-4614-4077-2. DOI 10.1007/978-1-4614-4078-9

[Vasquez 2012] VAZQUEZ, J., CHAMPAC, V., ZIESEMER, A., REIS, R., TEIXEIRA, I., SANTOS, M. and TEIXEIRA, P., Delay Sensing for Long-Term Variations and Defects Monitoring in Safety–Critical Applications, IN: Analog Integrated Circuits and Signal Processing, Volume 70, Number 2, 249-263, February 2012, Springer, ISSN 0925-1030, DOI: 10.1007/s10470-011-9789-0.

[Violante 2011] VIOLANTE, M., MEINHARDT, C., REIS, R., REORDA, M., A Low-Cost Solution for Deploying Processor Cores in Harsh Environments, IEEE Transactions on Industrial Electronics, Vol. 58, Issue 7, p. 2617- 2626, July 2011, ISSN: 0278-0046, DOI: 10.1109/TIE.2011.2134054.

PROCESSAMENTO DE CONSULTAS ESPACIAIS EM REDES DEPENDENTES DE TEMPO DE LARGA ESCALA

José A. F. de Macêdo[1], Regis P. Magalhães[1], Vania Vidal[1], Marco A. Casanova[2], Mario Nascimento[3], Samara M. Nascimento[1], Camila F. Costa[1], Raffaele Perego[4], Chiara Renso[4], Regis Melo[5]

Abstract. Applications related to urban mobility are key for helping individuals, businesses and government agencies in managing displacements within the urban space. Such applications usually support services for route planning, detection of shortest paths, queries of nearest neighbors and others. A common characteristic of these applications is that they employ a structure of the underlying road network. Indeed, the nature of such networks is spatio-temporal. They need to represent space and spatial constraints to which the moving objects are submitted. They also need to consider that the travel time on road networks depends directly on the traffic and that traffic patterns depend, in turn, not only on the region in question, but also the time of day. Therefore, the time a moving object takes to cross a path segment typically depends on the starting instant of time. Thus, the time taken by a moving object to traverse a route in a road network depends on the initial time that it started the route. So, we call time-dependent networks, the networks with this spatio-temporal feature. In this project we aim to address the problem of processing spatial queries in large-scale time-dependent networks, since there is a need for real world applications dealing with large-scale networks and large volumes of moving objects. Furthermore, we intend to develop a software platform that enables the development of applications using time-dependent networks. As proof of concept, we will use this platform on real problems of three companies that provide mobility services.

1Departamento de Computação, Universidade Federal do Ceará (DC/UFC) Campus do Pici, Bloco 910,60.455-760, Fortaleza, BRASIL
2Depto. de Informática, Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO) Ruas Marquês de São Vicente, 225, Rio de Janeiro, BRASIL
3Department of Computing Science, University of Alberta
Edmonton, Alberta, CANADÁ
4Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (KDDLab/HPLab/ISTI/CNR) via G. Moruzzi 1, 56124, Pisa, ITÁLIA
5Sagarana Tech
Avenida Engenheiro Luis Vieira, 920 sala 111, Fortaleza, BRASIL

{jose.macedo,regispires,vvidal,camilaferc,samara.martins}@lia.ufc.br, casanova@inf.puc-rio.br, nascimento@ualberta.ca, {raffaele.perego,chiara.renso}@isti.cnr.it, regismelo@sagaranatech.com

Resumo. *Aplicações relacionadas com mobilidade urbana são chave para ajudar pessoas, empresas e órgãos governamentais na gestão de deslocamentos no espaço urbano. Tais aplicações visam prover serviços relacionados com planejamento de rotas, detecção de menores caminhos, consultas de vizinhos mais próximos e outras. Algo comum entre essas aplicações é o fato delas utilizarem uma estrutura de rede de rodovias subjacente. A natureza de tais redes é espaço-temporal. A dimensão espacial advém do fato de que elas precisam representar o espaço e as restrições espaciais às quais os objetos móveis são submetidos. A dimensão temporal das redes ocorre porque o tempo de viagem em redes rodoviárias depende diretamente do tráfego e que os padrões de tráfego dependem, por sua vez, não apenas da região em questão, mas também do período do dia. Portanto, o tempo que um objeto móvel leva para atravessar um segmento da via depende tipicamente do instante de tempo de partida. Assim, o tempo gasto para percorrer uma rota em uma rede de rodovias depende do tempo inicial em que se iniciou o percurso. Denominamos redes com esta característica espaço-temporal de redes dependentes do tempo (RDT). Neste projeto, visamos atacar o problema de processamento de consultas espaciais em redes dependentes do tempo de larga escala, dada a necessidade das aplicações lidarem com redes de larga escala e com volumes grandes de objetos móveis. Além disso, pretendemos desenvolver uma plataforma de software que permita o desenvolvimento de aplicações que utilizem redes dependentes do tempo. Como prova de conceito, usaremos esta plataforma em problemas reais de três empresas que proveem serviços de mobilidade.*

1. DESCRIÇÃO DO PROBLEMA

A alta disponibilidade de dispositivos de rastreamento, tais como equipamentos habilitados com GPS, propiciou a expansão dos chamados serviços baseados em localização. Tais serviços levam em consideração a posição geográfica de uma entidade [Junglas e Watson 2008]. Esta mesma disponibilidade possibilita também coletar grande quantidade de dados de mobilidade de veículos, pessoas ou quaisquer outros objetos móveis. Uma série de diferentes e importantes análises podem ser feitas sobre dados de mobilidade, permitindo descobrir, por exemplo, o comportamento e as preferências de indivíduos. No contexto de trânsito, é possível coletar uma grande quantidade de dados de trajetórias de veículos e, a partir dos mesmos, criar uma visão das condições de tráfego no tempo e espaço. A avaliação das condições de tráfego é a chave para propiciar sistemas de transporte inteligentes.

Aplicações relacionadas com mobilidade urbana são chaves para ajudar pessoas, empresas e órgãos governamentais no gerenciamento de deslocamentos no dia-a-dia. Tais aplicações têm como objetivo planejamento de rotas, detecção de menores caminhos, consultas de vizinhos mais próximos e outras. Em geral tais aplicações utilizam uma estrutura de rede de rodovias subjacente. A natureza de tais redes é espaço-temporal.

Espacial, pois precisam representar o espaço e as restrições espaciais às quais os objetos móveis são submetidos. E temporal, considerando que o tempo de viagem em redes rodoviárias depende diretamente do tráfego e que os padrões de tráfego dependem, por sua vez, não apenas da região em questão, mas também do período do dia. Portanto, o tempo que um objeto móvel leva para atravessar um segmento da via depende tipicamente do instante de tempo de partida. Assim, o tempo gasto para percorrer uma rota em uma rede de rodovias depende do tempo inicial em que se iniciou o percurso. Essa variação temporal do tempo de viagem é fruto das características dinâmicas do sistema causadas, por exemplo, por engarrafamentos em alguns momentos do dia, e de pouco ou nenhum tráfego em certas regiões e horários. A Figura 1 (obtida do Google Maps[1]) apresenta um exemplo real de como o tempo pode influenciar no tráfego, portanto, no tempo gasto no percurso de uma viagem. Ela ilustra dois momentos diferentes de uma mesma via na cidade de Fortaleza. As 7h30min da manhã (Figura 1a), o tráfego é mais intenso (em vermelho) e, portanto, o tempo para percorrer a via é maior que ao meio-dia (Figura 1b), quando não há congestionamentos.

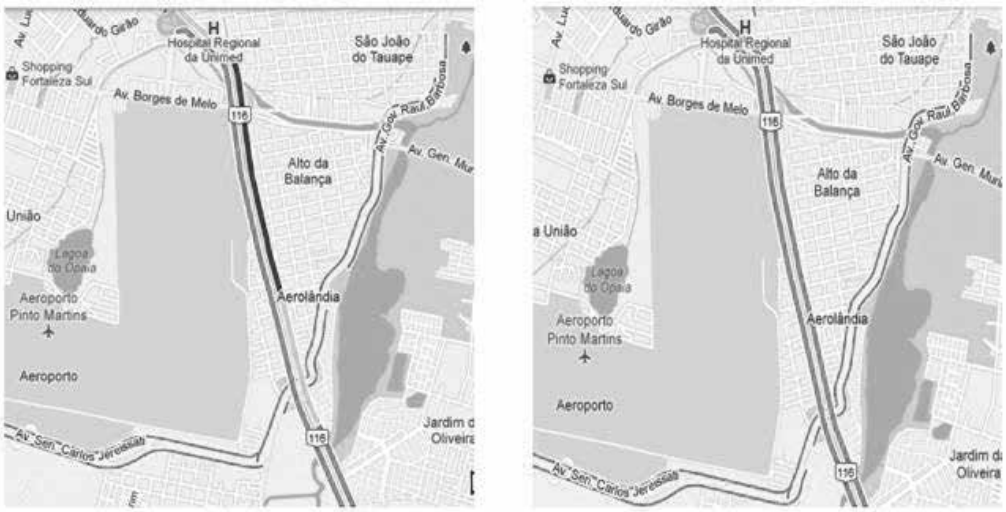


Figura 1. BR-116 em dois diferentes momentos. Do lado esquerdo (a), às 7h30min da manhã. As cores vermelha e amarela da via representam tráfego intenso. Do lado direito (b), às 12h30min, a cor verde representa trânsito livre na mesma via.

[1]http://maps.google.com/

Pesquisas recentes têm incluído a dependência temporal para solucionar problemas convencionais de consultas espaciais [Goldberg e Harrelson 2005][Nannicini 2012] [Demiryurek 2011][George et al. 2007]. Entretanto, a maioria dos trabalhos existentes que propõem soluções para consultas espaciais não consideram a variação temporal. Neles, a distância entre dois pontos na rede é fixa e independente do tempo, o que não reflete as condições de tráfego previstas em redes reais.

Desta maneira, um requisito fundamental para atender às diversas aplicações que precisem considerar a dependência temporal, é utilizar um modelo de rede onde a distância entre dois pontos da rede seja representada por uma função temporal. Estas funções fornecem o tempo de viagem de uma via da rede em função do instante de tempo em que se inicia um percurso. Desta forma, além das informações espaciais, é possível incorporar à rede as informações sobre os padrões de tráfego em cada via. Construir estratégias e algoritmos para o processamento correto e eficiente de consultas em redes dependentes do tempo é um desafio, desde que as propriedades comuns de grafos podem não ser satisfeitas no caso dependente do tempo [George et al. 2007].

Neste sentido, pesquisamos e desenvolvemos soluções para resolver consultas (e.g. kNN, kNN-TimeToService e OSR, discutidas na Seção Estado-da-Arte) sobre redes dependentes do tempo, publicados em [Cruz et al. 2012][Costa et al. 2013][Costa et al. 2014]. Particularmente, em [Cruz et al. 2012], apresentamos a implementação de uma estrutura de dados que permite a indexação de redes dependentes do tempo de forma eficiente. No entanto, apesar dos esforços realizados até o momento, percebemos que aplicações reais utilizam redes muito grandes, chegando à ordem de bilhões de nós e arestas, fato que se torna mais complexo quando tratamos de redes dependentes do tempo. Neste documento, chamaremos tais redes de Redes Dependentes do Tempo em Larga Escala (RDT-LE). Neste contexto, muitos desafios precisam ser superados ao lidarmos com RDT-LEs, tais como: (1) criação e atualização da rede, (2) armazenamento distribuído da rede, (3) implementação de métodos de acesso e (4) processamento eficiente de consultas.

Diante deste contexto, visamos neste projeto atacar algumas questões relacionadas com os quatro grandes desafios mencionados anteriormente. Além disso, assumiremos que para lidarmos com RDT-LE deveremos recorrer às infraestruturas de nuvens computacionais, as quais permitam usar o poder computacional de grandes *clusters* de computadores para escalar as soluções desenvolvidas. Neste contexto, tentaremos responder às seguintes perguntas de partida, que nos guiarão ao longo da nossa pesquisa:

Como criar RDT-LE a partir de dados de trajetórias de objetos moveis?

Qual é a melhor estratégia para particionar os dados de uma RDT-LE para permitir o processamento escalável de consultas?

Como processar eficientemente consultas do tipo KNN, OSR e junções espaciais sobre RDT-LEs?

Como devemos realizar a atualização de uma RDT-LE, minimizando o impacto da atulizacao no processamento de consultas sobre a rede?

Quais são as melhores estratégias para indexar uma RDT-LE?

Além dos objetivos de pesquisa, este projeto visa construir uma plataforma para processamento de consultas em RDT-LE. Esta plataforma visa prover uma infraestrutura básica para a implementação de algoritmos para criação, armazenamento, indexação e processamento de consultas em grandes redes dependentes do tempo.

2. ESCOPO DA SOLUÇÃO PROPOSTA

2.1. Redes Dependentes do Tempo

A estrutura de uma rede pode ser modelada por um grafo onde os vértices representam as junções, pontos iniciais e finais de um segmento de rodovia e as arestas conectam os vértices. O tempo de viagem é modelado por um grafo dependente do tempo, TDG (do inglês *Time-Dependent Graph*), onde o custo (tempo) para percorrer uma aresta é dado por uma função do instante de tempo de partida. O conceito de TDG é formalizado de acordo com as definições de *travel-time*, *time-dependent fastest path* e *time-dependent distance* apresentadas a seguir.

Definição 1. Um TDG $G = (V, E, C)$ é um grafo onde: (i) $V = \{v_1,...v_n\}$ é um conjunto de vértices; (ii) $E = \{(v_i,v_j) \mid v_i,v_j \in V, i \neq j\}$ é um conjunto de arestas; (iii) $C = \{c(v_i, v_j)(\cdot) \mid (v_i, v_j) \in E\}$, onde $c(v_i,v_j) : [0,T] \rightarrow R+$ é uma função que atribui um peso positivo para (v_i, v_j) dependendo de um instante de tempo $t \in [0, T]$ e onde T é o tamanho do domínio temporal.

Em termos gerais, um TDG é um grafo onde o custo das arestas varia com o tempo. Para cada aresta (u, v), uma função c(u, v)(t) dá o custo de percorrer a aresta quando o percurso é iniciado no instante de tempo t . O domínio das funções em C é representado como [0, T]. Por exemplo, quando T = 24h, qualquer função c(u, v)(t) está definida para um dia completo. A Definição 1 não restringe ao TDG a ser não direcionado, de tal forma que a existência de uma aresta (u, v) não implica na existência de (v, u). Além disso, ela permite que arestas opostas, (u, v) e (v, u), possam ser tais que $c(u, v)(t) \neq c(v, u)(t)$.

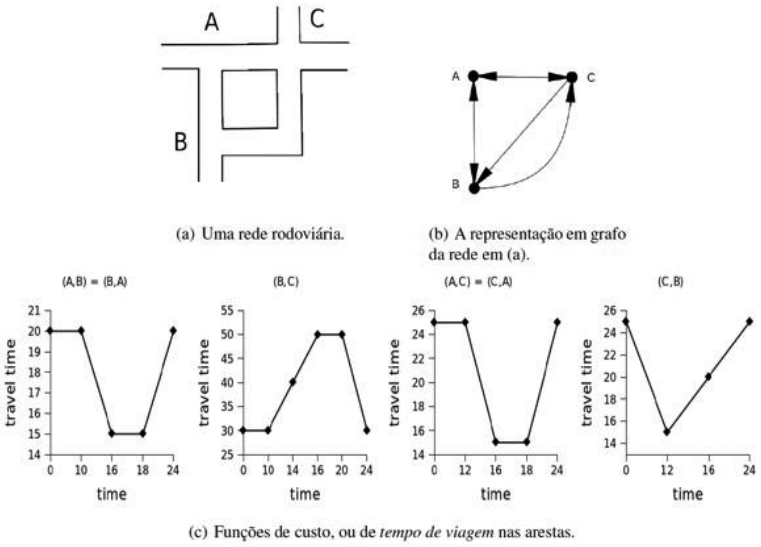


Figura 2. Exemplo de uma rede rodoviária, o grafo que modela a rede e as funções de tempo.

Para exemplificar, considere a Figura 2(a) onde uma rede rodoviária é parcialmente apresentada. A rede pode ser modelada pelo grafo apresentado na Figura 2(b). O tempo de percurso de cada aresta é dado pelas funções da Figura 2(c). Os pares de arestas opostas (B, A) e (A, B), e (A,C) e (C,A), têm o mesmo custo. Entretanto, (B,C) e (C,B), apesar de opostas, possuem custos distintos. O custo temporal para executar um caminho a partir de um instante de tempo específico é chamado de *travel-time*. O *travel-time* é calculado assumindo que paradas não são permitidas. Para seu cálculo deve-se calcular o tempo de chegada em cada aresta.

Dado que um veículo percorre um segmento de rodovia e este percurso inicia em um determinado instante de tempo t, o tempo de chegada é o instante de tempo no qual o veículo deve chegar ao outro extremo da via. Considerando a Figura 2(c), por exemplo, ao percorrer a via representada por (B;C) às 10h00min horas da manhã, o veículo chega às 10h30min. Para exemplificar, considere o caso em que t = 24:00, logo o custo é dado por c(B;C)(24:00) = 0:30, pois o veículo deverá chegar às 0h30min.

A definição acima trata de mostrar como o custo de uma rota, *travel-time*, deve ser calculado. A ideia é simples. Dada a sequência de arestas do grafo que representam a rota iniciada no instante t, o custo da primeira é calculado com respeito ao instante t. O custo das próximas arestas é dependente do *arrival-time* da aresta anterior. Note que essa definição de custo não leva em consideração paradas nos nós do grafo, ou seja, o percurso da próxima aresta do caminho inicia exatamente no mesmo instante de tempo em que esta foi alcançada.

Retornando ao exemplo das Figuras 2(b) e 2(c), considere dois caminhos para ir de B até C. Alguém poderia preferir o caminho B → C, que passa por B e vai diretamente ao C. Outra possibilidade é o caminho B→A→C, que passa por A. O caminho mais rápido de B a C depende do instante de partida t_s. Para t_s = 10:00, o tempo de viagem do caminho B→C é de 30 minutos.

O caminho B→A→C, iniciado às 10h00min tem o tempo de viagem calculado da seguinte forma. O custo de percorrer às 10:00 é c(B→A)(10:00) = 0:20, o tempo de chegada é 10h20min. O custo de percorrer a próxima aresta (A→C) é dado por c(A→C) (10:20) = 0:45. Portanto, o custo total é de 45 minutos. Similarmente, se t_s = 16:00, então o tempo de viagem de B→C é 50 minutos e de B→A→C é 30 minutos. Dessa forma, o caminho mais rápido às 10h00min é 20 minutos. No entanto, às 16 horas o caminho mais rápido leva 30 minutos para ser percorrido.

2.2. Problema do Menor Caminho (estático) e Problema do Caminho mais rápido (dependente do tempo)

Estes problemas estão intrinsecamente relacionados com a solução de consultas de TDk-NN, desde que seu resultado depende do custo (tempo de viagem) do mais rápido para cada ponto de interesse do resultado. A solução mais conhecida para o problema do menor caminho em redes estáticas (não dependente do tempo) é o algoritmo de Dijkstra [Dijkstra 1959]. Em sua execução os nós do grafo são visitados a partir do nó fonte, em ordem crescente da distância para o mesmo, até alcançar o nó destino. Nesta seção, serão descritas algumas técnicas que foram propostas para acelerar a execução do algoritmo de Dijkstra, bem como para a solução da versão dependente do tempo do problema de menor caminho.

Além do algoritmo de Dijkstra, muitas outras soluções para o problema de menor caminho foram propostas. A maior parte destas soluções é definida por técnicas que tem o objetivo de acelerar o tempo de execução do algoritmo de Dijkstra. Elas são em geral baseadas em etapas de pré-processamento que anotam a rede com informações utilizadas para podar ou guiar a busca. Em [Wagner e Willhalm 2007] pode-se encontrar um estudo sobre algumas das técnicas aplicadas a redes estáticas. Entretanto, vale ressaltar que as técnicas utilizadas em redes estáticas não são diretamente aplicáveis no caso dependente do tempo. Muito menos esforço foi dedicado ao caso em que a rede é dependente do tempo. O primeiro algoritmo que considera a variante dependente do tempo do problema de menor caminho é abordado em [Cooke e Halsey 1966]. Este algoritmo é uma forma modificada do algoritmo de Bellman para encontrar o caminho mais rápido entre dois vértices em uma rede. A seguir, são apresentadas algumas destas técnicas propostas para acelerar a execução do algoritmo de Dijkstra.

Uma das técnicas utilizadas para melhorar o desempenho do algoritmo de Dijkstra é a Busca Bidirecional. Um algoritmo baseado em busca bidirecional realiza simultaneamente duas buscas unidirecionais: uma busca na direção original (*forward*) e uma busca na direção contrária (*backward*). Na busca *forward* a árvore de busca cresce a partir do nó fonte na direção do nó destino, na qual a distância mínima da fonte para um conjunto de nos é descoberta. A busca *backward* acontece na direção reversa, do nó destino para o nó fonte. Nesta busca, a distância mínima de um conjunto de nós para o nó destino é descoberta. A busca *backward* é aplicada ao grafo reverso, ou seja, o grafo cujo conjunto de arestas é formado pelas arestas do grafo original na direção oposta. Dessa forma, por serem mantidas duas árvores de busca (*forward* ou *backward*) , uma para cada direção, uma das escolhas a ser feita ao utilizar uma busca bidirecional é qual arvore de busca será escolhida para expansão no próximo passo. Outra técnica é a Busca Direta ou Busca A∗. A busca direta utiliza conhecimentos do domínio do problema para alterar a prioridade dos vértices enfileirados e, portanto, a ordem em que estes são visitados. A ideia é dar maior prioridade aqueles que tem o maior potencial de alcançar o objetivo. No caso do problema de menor caminho, o objetivo é o vértice destino. O potencial é dado por uma função heurística que deve ser consistente.

Em [Goldberg e Harrelson 2005] é apresentada uma proposta baseada em busca direta para solucionar o problema do menor caminho em grafos estáticos. A solução proposta neste trabalho combina a busca direta com busca bidirecional e uma outra técnica baseada em marcos (em inglês, *landmarks*, termo utilizado na literatura) e na desigualdade triangular. Nesta estratégia, inicialmente $d = \infty$. Então, partindo de um vértice fonte s e um vértice destino t, quando uma aresta (v,w) é avaliada pela busca *forward* e w já foi avaliado pela direção reversa, os menores caminhos s – v e w – t já são conhecidos, com tamanhos df(v) e db(w). Sendo l(v,w) um limite inferior para a distância entre v e w, se $d < df(v) + l(v,w) + db(w)$, um caminho menor que o menor caminho visto até então foi encontrado, portanto o valor de d é atualizado. O algoritmo termina quando a busca em uma das direções encontra um vértice que foi avaliado na outra direção. Para estimar os limites inferiores l(v,w), a estratégia seleciona um conjunto de *landmarks* (vértices específicos da rede) e, para cada vértice, pré-computa d para todos os *landmarks*. Considerando um *landmark* L e seja d(.) uma distância para L. Então, pela desigualdade triangular, $d(v) – d(w) < dist(v,w)$. Similarmente, se d(.) é uma distância de um caminho partindo de L, $d(w) – d(v) < dist(w,v)$. O limite inferior escolhido é o máximo entre todos os limites inferiores calculados utilizando os *landmarks*.

Em [Nannicini 2012] é proposta uma solução para o problema do menor caminho em redes dependentes do tempo que utiliza também uma estratégia baseada em busca direta e busca bidirecional. A solução proposta visa reduzir o espaço de busca, entretanto, não alcança resultados ótimos. Em redes dependentes do tempo a busca bidirecional não pode ser aplicada diretamente, pois o tempo de chegada (*arrival-time*) no destino é

desconhecido, e este tempo influenciaria no custo da busca *backward*. Para solucionar este problema, a solução proposta utiliza na busca *backward* limites inferiores sobre custo dos arcos para restringir o conjunto de nós a ser explorado na busca forward. A busca direta é feita através do uso de *landmarks*. Em redes não-dependentes do tempo, como já foi exemplificado, *landmarks* podem ser usados na implementação da busca bidirecional. Para isso, a função heurística deve ser consistente para ambas as buscas, *forward* e *backward*. No método proposto em [Nannicini 2012], a busca *backward* é executada sobre o grafo de limites inferiores. Para cada aresta, o custo neste grafo é o limite inferior do custo da função na aresta. A busca *backward* é utilizada para fornecer limites para o critério de poda da busca *forward*. De maneira geral, a estratégia consiste em seguir com as duas buscas até que a busca *backward* contenha apenas os vértices cuja chave associada não exceda um parâmetro μ, onde μ é um limite superior do custo da solução ótima. Quando a busca *backward* para, todos os vértices avaliados por ela são incluídos em um conjunto M, de tal forma que a busca forward, a partir de então, avalia somente os vértices incluídos em M.

Em [Demiryurek 2011] é avaliada a computação on-line do caminho mais rápido em redes de rodovias dependentes do tempo e apresenta uma técnica baseada em busca bidirecional e A que acelera a computação do caminho. A estratégia particiona a rede em partições não sobrepostas. Uma pré-computação cria as partições e computa o limite inferior das distâncias entre as bordas das partições, entre vértices e borda das partições e entre bordas e vértices. O particionamento atribui a cada vértice um conjunto de partições. Uma outra etapa computa para cada par de partições o limite inferior do custo caminho mais rápido entre elas. Todos os custos das distâncias entre os vértices bordas das partições são armazenados. O algoritmo on-line visita todos os nós alcançáveis a partir da consulta em direção ao destino, até que o destino seja alcançado. A busca aplicada é direcionada, a função heurística é obtida a partir dos valores de limites inferiores da distância entre bordas das partições, obtidas na pré-computação.

2.3. Consultas dos K-Vizinhos mais próximos em Redes Dependentes do Tempo

Enquanto alguns trabalhos têm sido propostos para acelerar a computação dos menores caminhos em redes dependentes do tempo, poucos trabalhos recentes propõem métodos para processar consultas kNN nessas redes. Esta seção apresenta os trabalhos que propõem soluções para o processar consultas kNN em redes dependentes do tempo (TD-kNN, do inglês Time-Dependente k Nearest Neighbors).

Uma consulta TD-kNN considera a dependência temporal para responder consultas kNN. Uma consulta kNN recupera o conjunto de k pontos de interesse que são mais próximos a um ponto de consulta. Nas redes dependentes do tempo, uma consulta kNN retorna o conjunto de k pontos de interesse com o mínimo tempo de viagem do ponto de

consulta, considerando um instante de partida. Como exemplo, o cenário ilustrado na Figure 3. Imagine um turista em Paris, interessado em visitar alguma atração turística próxima a ele. Considere duas atrações turísticas na cidade, a Torre Eiffel e a Catedral de Notre Dame. O turista gostaria de saber qual destes pontos turísticos tem um caminho do ponto onde ele está em determinado instante, que é o mais rápido entre os possíveis caminhos considerando as condições de tráfego durante o percurso. Por exemplo, caso a consulta seja processada às 10 horas da manhã ele levaria 20 minutos para ir para catedral, sendo esta a atração mais próxima. Se ao invés disso, a consulta for submetida às 22 horas, no mesmo ponto de partida, a atração mais próxima seria a Torre Eiffel.

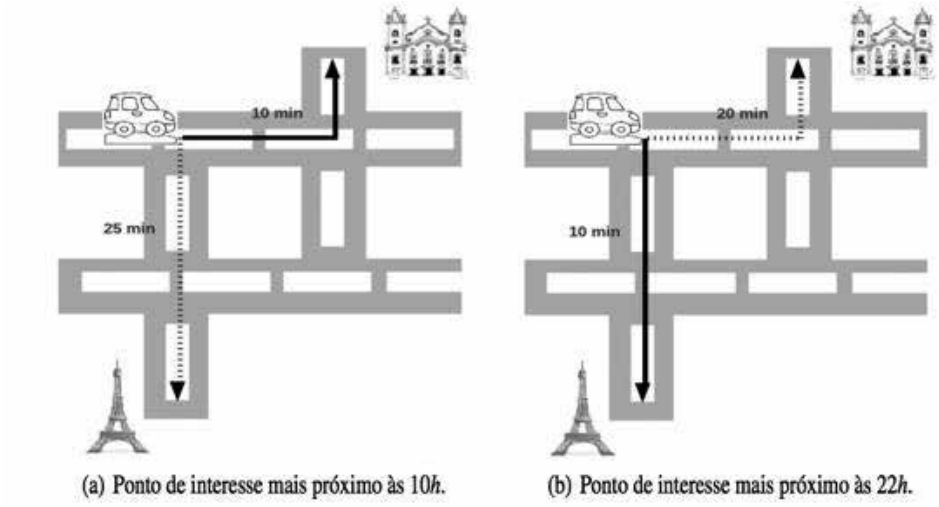


Figura 3. Exemplo de RDT

2.4. ALGORITMOS DE EXPANSÃO INCREMENTAL

O problema de solucionar consultas kNN em redes dependentes do tempo foi introduzido em [Demiryurek et al. 2010]. Neste trabalho, os autores comparam duas diferentes soluções para este problema. Os métodos apresentados são uma adaptação do algoritmo INE (*Incremental Network Expansion*) para redes dependentes do tempo. Uma das soluções aplica expansão incremental enquanto utiliza o modelo de rede expandida no tempo como modelo de representação da rede. Este modelo discretiza o domínio temporal $T = [t_0, t_n]$ em n instantes de tempo igualmente espaçados $t = 0, 1, ..., t_n$ e constrói uma rede estática para cada instante de tempo. Portanto, o modelo de rede expandida no tempo replica a rede original n vezes. Então, o modelo conecta cada vértice de um determinado instante em t aos seus vizinhos no próximo instante, calculado pelo instante atual mais o custo da aresta nesse instante. Embora este modelo possa utilizar

os mesmos algoritmos aplicados a redes estáticas, as soluções que o utilizam podem apresentar resultados incorretos. Neste modelo, o custo de uma aresta para todo um intervalo é dado pelo custo do instante inicial, resultando em um erro entre o valor do menor caminho e do caminho calculado.

O segundo método é o algoritmo TD-NE (*Time Dependent Network Expansion*) que expande os vértices da rede na ordem de menor distância para o ponto de consulta, considerando as funções de custo. Uma tabela *hash* que associa as funções de custo às arestas da rede é consultada durante a expansão para atualizar o custo dos caminhos avaliados.

2.5. CONSULTAS K-VIZINHOS MAIS PRÓXIMOS COM RESTRIÇÕES DE TEMPO DE OPERAÇÃO (KNN-TIME-TO-SERVICE)

Consultas dos k-Vizinhos mais próximos em RDT retornam os k pontos de interesse que são mais próximos em termos de tempo de viagem de ponto de consulta q, dado um determinado tempo de partida. Como apenas o tempo de viagem é considerado, a resposta para esta consulta pode levar a pontos de interesse que são próximos a q, mas não estão operacionais, ou seja, estão fechados para atendimento e podem levar um grande tempo para estarem abertos. Processar k-vizinhos mais próximos considerando o tempo de operação das pontos de interesse em um RDT é importante para aplicações reais. A solução para este problema, diferente das soluções anteriores é que precisamos minimizar o tempo para ser servido e não apenas para alcançar o ponto de interesse.

Propusemos em [Costa et al. 2014] três soluções para este problema. Cada solução requer diferentes formas de pré-processamento de forma a calcular os limites para guiar a busca de pontos de interesse, que são usados para realizar a poda na busca da solução. A decisão de propor três soluções distintas é justificada pelo fato de aplicações de mobilidade requisitarem diferentes taxas de atualização da rede. Por exemplo, em uma aplicação de gerenciamento de tráfego, as redes devem ser atualizadas com mais frequência, para refletir situações do trânsito, diferentemente de aplicações turísticas as quais requerem que atualização ocorra apenas quando o tempo de operação dos locais turísticos é alterada. Consequentemente, como o pré-processamento é um passo fundamental para nossa solução, decidimos prover três soluções para acomodar cenários com redes de baixa e alta taxa de atualização. A primeira solução usa limites relaxados visando reduzir o custo do pré-processamento. A segunda solução possui um pré-processamento intermediário, e usa limites mais rígidos que a primeira e mais relaxados que a última solução. Finalmente, a terceira solução, requer um pré-processamento mais custoso de forma a calcular limites mais restritos.

2.6. ÍNDICES TNI E LNI

Em [Demiryurek et al. 2010] é apresentada uma solução que assemelha-se com uma versão da solução baseada em células de Voronoi em redes dependentes do tempo. Nesse trabalho, é proposto um processo de pré-computação que constrói duas estruturas que indexam a rede e os pontos de interesse chamadas de *Tight Network Index* (TNI) e *Loose Network Index* (LNI). Ambas as estruturas são compostas por células que referenciam um ponto de interesse em particular, tal que, se um ponto de consulta q está dentro de uma célula de um ponto de interesse p no TNI, então certamente p é seu vizinho mais próximo. Entretanto, se q está fora de uma célula de um ponto de interesse p no LNI, então p certamente não é o vizinho mais próximo de q. Como na estratégia NVP para redes estáticas, usando o TNI pode-se encontrar imediatamente o primeiro ponto de interesse mais próximo ao objeto de consulta. Para os demais pontos de interesse ($k > 1$), o vizinho mais próximo seguinte é o ponto de interesse gerador de uma das células vizinhas às células dos pontos de interesse já encontrados. Para decidir qual o próximo ponto de interesse, a rede é expandida incrementalmente até encontrar um ponto de interesse em uma das células vizinhas.

2.7. CONSULTA ROTA SEQUENCIADA ÓTIMA (OPTIMAL SEQUENCED ROUTE - OSR)

A consulta de rota sequenciada (em inglês *Optimal Sequenced Route* – OSR) foi introduzida por [Kolahdouzan e Shahabi 2005] . Esta consulta foi utilizada em várias aplicações para planejamento de viagens, tais como, serviço baseado em localização e sistemas de navegação automotivo. Esta consulta objetiva achar a rota ótima a partir de uma localização de origem para uma localização de destino, passando por um certo numero de pontos de interesse em uma determinada sequência de acordo com as categorias desses pontos de interesse. Por exemplo, suponha que uma pessoa queira sair de casa em direção a um banco para tirar dinheiro, e então deseje visitar um shopping center para comprar roupas antes de retornar para casa. A restrição que determina que ela deva primeiro passar no banco é porque precisa ter dinheiro para comprar mercadorias no shopping. Embora existam muitos bancos e *Shopping Centers* na cidade, a consulta OSR escolhe o banco e o *Shopping Center*, na ordem definida, minimizando o custo total da viagem.

Algumas soluções para melhoria da consulta OSR foram propostas na literatura [Sharifzadeh e Shahabi 2008][Ohsawa et al. 2012][Htoo et al. 2012][Eisner e Funke 2012]. A consulta endereçada nesses trabalhos não levam em consideração que, tipicamente, o tempo que um objeto móvel leva para atravessar um segmento depende do tempo de partida. Diferentemente dos trabalhos anteriores, desenvolvemos uma solução denominada (TD-OSR) [Costa et al. 2013], a qual resolve consultas OSR sobre redes dependentes do tempo. Em tais redes, uma consulta OSR retorna a rota como o tempo mínimo

de viagem a partir de um ponto de origem para um ponto de destino considerando um determinado tempo de partida. Como estamos trabalhando com rede dependente do tempo, precisamos saber quanto tempo o usuário espera ficar em cada ponto de interesse, porque o tempo que a pessoa parte de um ponto de interesse afeta o resultado da consulta, visto que o tempo de viagem depende da hora da partida.

Em nossa solução foi utilizada como base uma estratégia de expansão incremental da rede, denominada INE [Papadias et al. 2003] e usou uma busca A* [Hart et al. 1968] para guiar a expansão, ou seja, para determinar a ordem na qual os vértices são expandidos na árvore de busca. Vértices com mais potencial são avaliados primeiro. Para medir o potencial de um vértice, este algoritmo usa a distância corrente do ponto da consulta para um dado vértice mais a função heurística, a qual no nosso caso é uma estimativa do tempo para alcançar as categorias de POIs pertencentes à sequência e o destino da consulta. Quanto menor o valor dado pela soma da distância corrente a um vértice mais sua função heurística, maior será seu potencial.

2.8. MÉTODOS DE ACESSO PARA REDES DEPENDENTES DO TEMPO

Construir métodos e técnicas para processar corretamente e eficientemente consultas sobre uma rede dependente do tempo é um desafio, dado que propriedades de grafos comuns não podem ser satisfeitas no caso de grafo dependente do tempo [George et al. 2007]. Particularmente, essas redes não podem ser armazenadas da mesma forma que uma rede estática. O mesmo se aplica para acessar a informação da rede. Consequentemente, este fato requer novos métodos para armazenamento de RDT que facilitem o acesso a informação da rede e suportem o projeto de algoritmos eficientes para implementar consultas sobre essas redes.

Algumas características de RDT devem ser consideradas no desenvolvimento deste método. Primeiramente, essas redes requerem mais espaço que as redes estáticas, pois necessitam armazenar os custos temporais para cada aresta. Desta forma, é preciso tentar manter o custo constante para atravessar uma aresta em cada intervalo de tempo. Outra observação importante é que o custo de armazenar as arestas de uma rede cresce de acordo com a granularidade dos intervalos de tempo. Além disso, armazenar em uma mesma página de disco todos os intervalos relativos ao custo de uma aresta, implica em acessar desnecessariamente dados que não serão usados, visto que as consultas visam buscar apenas um dado intervalo de tempo, de acordo com o tempo de partida do nó origem.

Baseados nessas informações, e objetivando processar consultas sobre RDT de forma eficiente, desenvolvemos um método para indexar um RDT, publicado em [Cruz et al. 2012]. O índice criado é composto por três níveis, o *Time-Level*, o *Graph-Level* e o *Data-Level* apresentado na Figura 4. As páginas de dados no *Time-Level* contêm ponteiros

para estruturas de índice no *Graph-Level*. Como uma RDT pode ser vista como um conjunto de grafos estáticos para cada intervalo de tempo, a ideia do primeiro nível é permitir acessar o primeiro grafo correspondente a um dado intervalo de tempo, evitando recuperar custos de arestas para todo tempo de partida possível. O *Graph-Level* tem uma estrutura de índice contendo ponteiros para uma página em disco no *Data-Level*, a qual armazena uma lista de adjacência dos vértices.

Uma arvore B+ é usada para armazenar a estrutura de índices para os níveis de *Graph* e *Time*, a qual foi provida pela biblioteca XXL [Bercken et al. 2001]. Os ponteiros para os níveis *Graph* e *Data* são armazenados nas folhas e cada nó (incluindo nós internos) é uma pagina em disco. Desta forma, o número de páginas acessadas para cada entrada de dado recuperada é da ordem de O(logb |P| + logb |V|), onde b é da ordem da árvore, e P é o numero de partições temporais que compõem o custo de uma aresta e V é o numero de vértices.

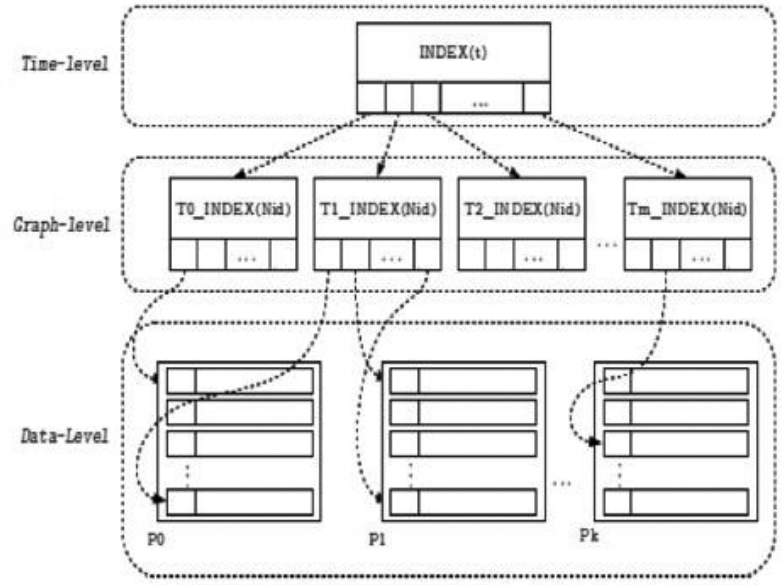


Figura 4. Índice para RDT

2.9. ARMAZENAMENTO E PROCESSAMENTO DE RDT EM LARGA ESCALA

Todas as atividades descritas acima podem ser realizadas num modelo de computação centralizado típico. Porém, é natural reconsiderá-los no domínio da computação em nuvem. Por exemplo, um possível caminho de pesquisa é a paralelização massiva dos índices propostos, e.g., usando o paradigma Map-Reduce, e.g. [Cary et al. 2009].

Além do processamento, o armazenamento em nuvem é outra questão a ser considerada. Não podemos ignorar que, subjacente ao desenvolvimento de todas atividades acima mencionadas, devemos explicitamente abordar o fato de que conjuntos de dados de trajetórias grandes e dinâmicos necessitarão ser armazenados, assim como manipulados para facilitar as tarefas de agrupamento (mineração) e indexação. Contudo, podemos vislumbrar inicialmente a maioria do processamento sendo feito de modo *off-line*, uma necessidade prática do sistema é estar habilitado para tratar fluxos de dados para melhor detectar padrões e/ou eventos de interesse em tempo hábil e de maneira distribuída.

3. Contribuições Pretendidas

O projeto em questão endereça a questão básica de execução de consultas em redes dependentes de tempo. Neste contexto, visamos propor teorias, técnicas e métodos para processar eficientemente tais consultas em RDT-LE. A aplicação dessas técnicas em um estudo de caso real é fundamental para a validação das abordagens propostas.

Os principais resultados esperados ao longo da execução do projeto são:

• A criação de uma infraestrutura para armazenar RDT-LE, a qual possa ser utilizada para processar consultas do tipo KNN, OSR e junções espaciais;

• Novos algoritmos para criar RDT-LEs a partir de trajetórias de objetos móveis;

• Métodos e técnicas para atualizar eficientemente RDT-LEs em tempo real;

• Novas estruturas de acesso para facilitar consultas espaciais sobre RDT-LEs;

• Desenvolvimento de algoritmos de consultas em paralelo para executar em infraestrutura de nuvem;

• Elaboração de trabalhos de iniciação científica e dissertações de mestrado e doutorado;

• Publicação de artigos em revistas e conferências nacionais e internacionais, indexadas pelo QUALIS;

• Prototipação de uma plataforma para processamento de consultas sobre RDT-LEs que contemple os algoritmos e as soluções propostas utilizando software livre. Assim será possível demonstrar os resultados também na forma de software em congressos nacionais e internacionais, em sessões de Demos. Esta etapa é importantíssima para que a indústria e outras instituições acadêmicas possam utilizar, na prática, os algoritmos e técnicas resultantes deste projeto de pesquisa.

• Finalmente, é esperado, como resultado deste projeto, a fortificação do grupo de pesquisa nesta área no Departamento de Computação da UFC e a integração deste grupo com os grupos de pesquisas parceiros desde projeto.

• Os principais indicadores de avaliação do projeto são:

• Publicação conjunta de 4 artigos em revistas e 6 artigos em congressos.

• Orientação, no Brasil, de 4 alunos de mestrado e 2 de doutorado, no tema do projeto.

• Protótipo de ferramenta com os algoritmos e soluções propostas, em software livre.

• Quantidade de módulos implementados na plataforma de processamento de consultas sobre RDT-LE.

3.1. Potencial de Inovação Tecnológica

A plataforma proposta por este projeto permitirá o desenvolvimento de inúmeros serviços de mobilidade para diversas áreas de aplicação, permitindo a criação de redes, seu processamento e recuperação de rotas dependentes do tempo. O serviços criados poderão ser aplicados em diversos segmentos, tais como, agências governamentais e empresa, as quais necessitem fazer a gestão da mobilidade de objetos móveis. O potencial de inovação será comprovado através de uma prova de conceito a ser realizada com as empresa SagaranaTech, TaxiSimples e Azul Tecnologia. Além disso, não existe no mercado produto gratuito que disponibilize um mapa com informações dependentes do tempo. Desta forma, esta infraestrutura será única e com grande potencial de uso.

4. PLANO PRELIMINAR PARA DESENVOLVIMENTO DA SOLUÇÃO

O objetivo principal deste projeto é desenvolver métodos e técnicas para processamento de consultas em RDT-LS. Além disso, este projeto visa desenvolver uma nova plataforma de software, de código aberto e livre, que possa ser utilizada para realizar e/ou desenvolver consultas sobre RDT-LS. Para cada um desses objetivos principais, destacamos os subobjetivos relacionados aos tópicos de pesquisa que deverão ser atacados ao longo deste projeto. Organizamos esses subobjetivos de acordo com as questões de pesquisa, destacando os tópicos específicos de pesquisa a serem tratados no contexto de cada subobjetivo.

4.1. Criação e atualização de RDT-LE

Iniciaremos as nossas pesquisas investigando métodos e técnicas para construção de RDT-LE utilizando trajetórias de objetos móveis que deverão ser utilizadas para compu-

tar as funções temporais que deverão anotar as arestas que ligam dois pontos da rede de ruas. Para que esta computação seja possível, precisaremos realizar uma operação denominada de *map-matching*, a qual deverá projetar os pontos das trajetórias sobre o arruamento, este processo é necessário pois em geral os dispositivos de GPS apresentam erros na localização exata dos objetos moveis.

Além da criação da RDT-LE, outro ponto muito importante a ser estudado é como atualizar um RDT-LE de forma incremental, evitando assim a sua completa reconstrução, o que tornaria inviável seu uso por aplicações do mundo real. A solução a ser projetada será bastante importante para resolver o problema de atualização da rede na ocorrência de problemas na rede de ruas, por exemplo, acidentes, obras, manifestações, etc. A empresa SagaranaTech tem um especial interesse por este problema, pois o seus algoritmos de replanejamento precisam ter informação recente sobre a mobilidade das vias para prever o horário de entrega das mercadorias e poder prover um replanejamento de rota eficiente.

4.2. Armazenamento distribuído de RDT-LE

Após o cálculo das funções temporais, será necessário um método para armazenar a rede de forma distribuída, visto que assumimos que o tamanho da rede é muito grande para ser armazenado e processado de forma centralizada. Neste sentido, pesquisaremos métodos e técnicas para particionamento de grandes grafos, visto que uma rede dependente do tempo será representada como um grafo direcionado. Uma questão importante a ser analisada é como realizar este particionamento levando em consideração a distribuição dos dados e os tipos de consultas a serem realizadas pelo grafo. Outra questão de pesquisa é como levar em consideração o espaço e tempo durante o particionamento.

Fará parte da pesquisa da distribuição dos dados de uma RDT-LT, a análise do uso de réplicas para aumentar a disponibilidade dos dados. Neste sentido, precisaremos avaliar as vantagens e desvantagens desta distribuição frente ao teorema CAP[2].

4.3. Processamento eficiente de consultas sobre RDT-LE

O processamento eficiente de consultas sobre um RDT-LE deve se apoiar em métodos e técnicas para processamento de grandes grafos. Neste sentido, investigaremos as estratégias recentes para processamento de grafos em larga escala, tais como, Pregel [Malewicz et al. 2010], Giraph [The Apache Project 2014], GraphLab [Low at al. 2010], GPS [Salihoglu e Widom 2013] e GraphX [Xin et al. 2014]. Claramente, tais estratégias

[2]http://en.wikipedia.org/wiki/CAP_theorem

são genéricas para qualquer tipo de grafo, no entanto nosso foco de pesquisa será avaliar e propor técnicas especificas para processamento de RDT-LE.

Com relação aos tipos de consultas a serem processadas, focaremos nossa pesquisa em consultas espaciais que desenvolvemos para RDT centralizadas [Cruz et al. 2012][Costa et al. 2013][Costa et al. 2014], tais como: TDNk-NN, TDNk-NN Time-to-Service, OSR. Além disso, ampliaremos essas consultas com novos tipos de consultas para atender ao problema de cobertura de trajetórias da empresa Azul Tecnologia e para consulta de descoberta do táxi mais próximo no tempo dada uma solicitação de táxi. Este último problema é de especial interesse da empresa TaxiSimples.

4.4. Implementação de métodos de acesso para RDT-LE

Para acelerar o processamento de consultas, deveremos prover métodos e técnicas que permitam acesso eficiente aos dados de uma RDT-LE. Com este foco, iremos estender a nossa abordagem de indexação de RDT para um ambiente distribuído. Além disso, estudaremos o impacto da distribuição dos dados neste mecanismo de indexação.

4.5. Desenvolvimento de Plataforma

Neste projeto temos um outro objetivo além dos resultados científicos, visamos o desenvolvimento de uma plataforma para processamento de consultas sobre RDT-LE. De fato, esta plataforma já vem sendo desenvolvida no contexto das dissertações e teses em andamento. No entanto, pretendemos com este projeto acelerar o seu desenvolvimento e viabilizar que tal plataforma seja disponibilizada de forma livre e aberta. Desta maneira, auxiliaremos novas pesquisas relacionadas ao processamento de consultas em RDT-LE. Claramente, esta plataforma será uma importante contribuição para a comunidade científica, bem como para o desenvolvimento de aplicações ligadas a redes dependentes do tempo.

A Figura 5 apresenta uma visão geral da plataforma que estamos desenvolvendo. As principais camadas da plataforma são: Armazenamento e Acesso à RDT-LE, Processamento de Consultas e Camada de Serviços. A camada de armazenamento e acesso à RDT-LE deverá permitir criação e armazenamento de redes utilizando dados de trajetórias de objetos móveis e redes de ruas. O armazenamento e acesso deverão recorrer à uma nuvem computacional para garantir a escalabilidade no processamento de grandes redes. A camada de processamento de consultas irá interpretar as consultas submetidas à plataforma e coordenar sua execução. A camada de serviços irá prover as consultas básicas sobre uma RDT, podendo ser estendida para novos tipos de consultas. Esta camada dará suporte à construção de aplicações no domínio de mobilidade.

Figura 5. Plataforma para Processamento de Consultas sobre RDT-LE

No atual momento, construimos a infraestrutura básica de classes e estamos implementando as consultas espaciais desenvolvidas nesta arquitetura.

5. CONCLUSÃO

Neste projeto visamos atacar o desafio de processar consultas espaciais sobre redes dependentes de tempo de larga escala. Como destacado neste documento, os grafos temporais de nosso interesse são aqueles utilizados para representar a mobilidade de objetos móveis sobre uma rede de vias. Tais redes temporais são de suma importância para prover soluções eficientes para problemas de caminho mínimo, vizinhos mais próximos, sequência ótima de roteamento, entre outros.

O problema relacionado com o processamento de grandes grafos vem sendo bastante estudado na literatura, motivado principalmente pelo volume crescente de dados que podem ser representados usando o modelo de grafos, tais como redes sociais e dados ligados, citando apenas alguns. No entanto, quando adicionamos a informação temporal a um grande grafo, passamos a lidar com dados com escalas de tamanho incrivelmente maiores. Neste contexto, pretendemos neste projeto desenvolver teorias, técnicas e métodos que nos permitam lidar com grafos temporizados de larga escala, visando principalmente processamento eficiente e eficaz.

REFERÊNCIAS

Bercken, J. V. D. et al. (2001). Xxl - a library approach to supporting efficient implementations of advanced database queries. In: Proc. of the 27th VLDB Conf. [S.l.: s.n.]. p. 39–48.

Cary, A. et al. (2009). Experiences on Processing Spatial Data with MapReduce. SSDBM 2009: 302-319.

Cooke, K. L.; Halsey, E. (1966). The shortest route through a network with time-dependent internodal transit times. Journal of Mathematical Analysis and Applications, v. 14, n. 3, p. 493–498. ISSN 0022-247X. Disponível em: <http://www.sciencedirect.com/science/article/pii/0022247X66900096>.

Costa, C. F.; Nascimento, M. A.; Macêdo, J. A. F.; Machado, J. C. (2013). Optimal Sequenced Route Queries in Time-Dependent Road Networks In: Proc. of the 2nd ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. p. 22–29.

Costa, C. F.; Nascimento, M. A.; Macêdo, J. A. F.; Machado, J. De C. (2014). A*- based Solutions for KNN Queries with Operating Time Constraints in Time-Dependent Road Networks. In: 15th IEEE International Conference on Mobile Data Management.

Cruz, L. A.; Nascimento, M. A., Macêdo, J. A. F. (2012). k-Nearest Neighbors Queries in Time-Dependent Road Networks. JIDM 3(3): 211-226.

Demiryurek, U. (2011). Online computation of fastest path in time-dependent spatial networks. In: PFOSER, D. et al. (Ed.). Advances in Spatial and Temporal Databases. Springer Berlin Heidelberg, (Lecture Notes in Computer Science, v. 6849). p. 92–111. ISBN 978-3-642-22921-3. Disponível em: <http://dx.doi.org/10.1007/978-3-642-22922-0_7>.

Demiryurek, U., Banaei-kashani, F., Shahabi, C. (2010) Efficient k-nearest neighbor search in time-dependent spatial networks. In: SPRINGER. Database and Expert Systems Applications. [S.l.]. p. 432–449.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. Numerische Mathematik, Springer-Verlag, v. 1, p. 269–271. ISSN 0029-599X. Disponível em: <http://dx.doi.org/10.1007/BF01386390>.

Eisner, J.; Funke, S. (2012). Sequenced route queries: Getting things done on the way back home. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems. [S.l.: s.n.]. (SIGSPATIAL '12), p. 502–505.

George, B.; Kim, S.; Shekhar, S. (2007). Spatio-temporal network databases and routing algorithms: a summary of results. In: Proceedings of the 10th International Conference on Advances in Spatial and Temporal Databases. Berlin, Heidelberg: [s.n.]. p. 460–477.

Goldberg, A. V.; Harrelson, C. (2005). Computing the shortest path: A* search meets graph theory. In: Proceedings of the Sixteenth Snnual ACM-SIAM Symposium on Discrete algorithms. Philadelphia, PA, USA: [s.n.]. p. 156–165.

Hart, P.; Nilsson, N.; Raphael, B. (1968). A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics, p. 100–107.

Htoo, H. et al. (2012). Optimal sequenced route query algorithm using visited poi graph. In: Web-Age Information Management. [S.l.: s.n.], (Lecture Notes in Computer Science, v. 7418). p. 198–209

Junglas, I. A., Watson, R. T. (2008). Location-based services. Commun. ACM, ACM, New York, NY, USA, v. 51, n. 3, p. 65–69, mar. 2008. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/1325555.1325568>.

Kolahdouzan, M., Shahabi, C. (2005). Alternative solutions for continuous k nearest neighbor queries in spatial network databases. GeoInformatica, p. 321–341.

Kriegel, H.-P. et al. (2007). Proximity queries in large traffic networks. In: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems. New York, NY, USA: ACM. (GIS '07), p. 21:1–21:8. ISBN 978-1-59593-914-2. Disponível em: <http://doi.acm.org/10.1145/1341012.1341040>.

Kriegel, H.-P. et al. (2011). Proximity queries in time-dependent traffic networks using graph embeddings. In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science. New York, NY, USA: ACM. (CTS '11), p. 45–54. ISBN 978-1-4503-1034-5. Disponível em: <http://doi.acm.org/10.1145/2068984.2068993>.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., Hellerstein, J. M (2010). Graphlab: A new framework for parallel machine learning. In UAI, pages 340–349.

Malewicz, G., Austern, M. H., Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., Czajkowski, G. (2010). Pregel: a system for large-scale graph processing. In SIGMOD Conference, pages 135–146.

Nannicini, G. (2012). Bidirectional A* search on time-dependent road networks. Networks, Wiley Subscription Services, Inc., A Wiley Company, v. 59, n. 2, p. 240–251. ISSN 1097-0037. Disponível em: <http://dx.doi.org/10.1002/net.20438>.

Ohsawa, Y. et al. (2012). Sequenced route query in road network distance based on incremental euclidean restriction. In: Database and Expert Systems Applications. [S.l.: s.n.], (Lecture Notes in Computer Science, v. 7446). p. 484–491.

Papadias, D. et al. (2003). Query processing in spatial network databases. In: Proc. of the 29th VLDB Conf. [S.l.: s.n.]. p. 802–813. ISBN 0-12-722442-4.

Salihoglu, S., Widom, J. (2013). GPS: a graph processing system. In Proceedings of the 25th International Conference on Scientific and Statistical Database Management (SSDBM). ACM, New York, NY, USA. Disponível em: <http://doi.acm.org/10.1145/2484838.2484843>

Sharifzadeh, M.; Shahabi, C. (2008). Processing optimal sequenced route queries using voronoi diagrams. GeoInformatica, v. 12, n. 4, p. 411–433.

The Apache Project. (2014). Apache Giraph, http://giraph.apache.org/. Visitado em: 13/05/2014.

Wagner, D.; Willhalm, T. (2007) Speed-up techniques for shortest-path computations. In: Proceedings of the 24th Annual Conference on Theoretical Aspects of Computer Science. Berlin, Heidelberg: [s.n.]. p. 23–36.

Xin, R. S., Crankshaw, D., Dave, A., Gonzalez, J. E., Franklin, M. J., Stoica, I. (2014). Graphx: Unifying data-parallel and graph-parallel analytics. CoRR, abs/1402.2394. Disponível em: <http://arxiv.org/abs/1402.2394>

REDES DE SENSORIAMENTO PARTICIPATIVO

Antonio A. F. Loureiro[1], Felipe França[2], Priscila M. V. Lima,[3], Leonardo B. Oliveira[1], Pedro Olmo S. Vaz de Melo[1], Thiago H. Silva[1], Olga N. Goussevskaia[1], Italo F. S. Cunha[1]

INTRODUÇÃO

No começo, havia *mainframes* compartilhados por um grande número de pessoas. Depois surgiu a era da computação pessoal, quando uma pessoa interagia diretamente com um computador e tinha total controle sobre ele. Depois, todos esses computadores e ou- tros dispositivos computacionais passaram a se conectar à Internet e o acesso a dados, aplicações e serviços passou a ser quase que instantâneo.

Atualmente, estamos presenciando a era da computação pervasiva, onde as Tec- nologias da Informação e Comunicação (TICs) estão sendo amplamente usadas, cada vez mais por mais pessoas, principalmente na forma de smartphones e tablets. Esses dispositi- vos, além de permitirem o acesso a informações, aplicações e serviços em qualquer lugar e a qualquer momento, permitem coletar dados do ambiente onde a pessoa se encontra de forma única, como nunca havia ocorrido. Ou seja, o espectro de sensoriamento de dados que está sendo disponibilizado através da Internet é extremamente amplo, incluindo desde as redes de sensores sem fio tradicionais (projetadas para aplicações como monitoramento ambiental e agricultura) até o sensoriamento social, onde o usuário tem um papel funda- mental na coleta de dados. As redes de sensores, onde as pessoas têm um papel decisivo na coleta de dados, são chamadas de Redes de Sensoriamento Participativo (RSPs, do inglês "*Participatory Sensing Networks*") e, no contexto, deste documento englobam as redes de sensores tradicionais, já que diferentes tipos de elementos de sensoriamento passarão a integrar e colaborar no processo de coleta de dados à medida que algoritmos mais sofis- ticados forem executados por esses elementos [Boukerche et al. 2014, Silva et al. 2014]. As RSPs possuem uma riqueza de dados coletados, o que abre possibilidades de desen- volvimento científico e tecnológico que estamos apenas começando a vislumbrar.

DISPOSITIVOS MÓVEIS E SENSORES

Elementos de sensoriamento estão se tornando ubíquos e podem ser encontrados em mui- tos dispositivos que fazem parte da solução de hardware para diferentes aplica-

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) Belo Horizonte, MG
[2]COPPE – Universidade Federal do Rio de Janeiro (UFRJ) Rio de Janeiro, RJ
[3]Núcleo de Computação Eletrônica – Universidade Federal do Rio de Janeiro (UFRJ) Rio de Janeiro, RJ

{loureiro,leonardo.barbosa,olmo,thiagohs,olga,cunha}@dcc.ufmg.br felipe@cos.ufpe.br, priscilamvl@gmail.com

ções, como a indústria automobilística, monitoramento do clima, saúde, controle industrial, explo- ração de petróleo, infra-estrutura de redes inteligentes, sistema de transporte inteligente e casas inteligentes, apenas para mencionar algumas. A rápida disseminação de senso- res em todos os domínios da vida humana se deve, principalmente, à tecnologia MEMS (*microelectromechanical systems*[1] ou sistemas microeletromecânicos[2]), que atingiu um nível de maturidade e está se tornando de uso comum. No futuro, podemos esperar ter mais e mais produtos baseados na tecnologia MEMS, que por sua vez permitirão o uso de mais sensores e o surgimento de novas aplicações.

Atualmente, os *smartphones* estão sendo fabricados com um número crescente de poderosos sensores embutidos de diferentes categorias, como acústica, som, vibração (por exemplo, microfone), corrente elétrica, potencial elétrico, magnético, rádio (por exemplo, magnetômetro), navegação (por exemplo, bússola, giroscópio), óptica, luz, imagem (por exemplo, luz ambiente, iluminação traseira, câmera), posição, ângulo, deslocamento, dis- tância, velocidade, aceleração (por exemplo, acelerômetro, GPS), pressão (por exemplo, barômetro) e proximidade (por exemplo, proximidade, toque). Esses são exemplos de sensores atualmente disponíveis em *smartphones*. De fato, *smartphones* topo de linha já possuem aproximadamente duas dezenas de sensores e possivelmente terão mais no futuro. Outros sensores, de diferentes categorias, podem ser facilmente incorporados a *smartphones* como da química (por exemplo, dióxido de carbono e monóxido de car- bono) e da área térmica, calor (por exemplo, temperatura). Sensores são extremamente importantes para desenvolver novas funcionalidades inteligentes para todos os tipos de dispositivos móveis.

Mais sensores em *smartphones* e outros dispositivos móveis significam que mais dados podem fluir por esses dispositivos e, eventualmente, para a nuvem com sua imensa capacidade de armazenamento e processamento. Isso também significa que processá-los para obter novas informações e conhecimento é fundamental. Atualmente, os *smartphones* já são utilizados por muitas aplicações de sensoriamento pessoal [Lane et al. 2010], tais como o acompanhamento de algum exercício físico (por exemplo, correr, caminhar). Nesse caso, o dispositivo é suficiente para proporcionar ao usuário a informação desejada, desde que o *smartphone* tenha os sensores incorporados e do software correspondente.

Aplicações de sensoriamento de multidões (*crowd sensing applications*) têm como objetivo monitorar fenômenos em grande escala e requerem a participação ativa das pes- soas e a disponibilização de seus dados coletados [Lane et al. 2010]. Nós prevemos que as aplicações de sensoriamento social serão fundamentais para a interação humana no futuro, nos tornando cientes para diferentes aspectos de nossas vidas. Há tantas oportuni- dades interessantes e grandes desafios à nossa frente na implantação de serviços móveis considerando esses dispositivos [Wang et al. 2013a, Wang et al. 2013b].

ESPECTRO DE SENSORIAMENTO

Sensores como acelerômetro 3D, giroscópio 3D e magnetômetro 3D estão se tornando padrão em dispositivos móveis para fornecer aos usuários aplicações mais sofisticadas. O sensor acelerômetro mede as variáveis de movimento linear x, y e z, o sensor giroscópio a inclinação, rotação e ângulos de rotação e o sensor magnetômetro o campo magné- tico nos eixos x, y e z. Cada um desses sensores tem recursos poderosos, mas também apresenta algumas limitações que afetam sua utilizasão em aplicações. Por exemplo, os acelerômetros são sensíveis a vibrações e podem gerar um sinal mesmo quando os dispo- sitivos móveis estão em repouso, giroscópios sofrem do *zero bias drift* (geram um sinal de saída do giroscópio quando não está ocorrendo qualquer rotação) e magnetômetros são sensíveis a interferências magnéticas e também geram um sinal indesejado.

Os sinais destes três sensores podem ser amostrados e enviados diretamente para diferentes aplicações, como um pedômetro (aplicativo para contar passos de uma pessoa ao andar). Neste tipo de aplicação, o objectivo é receber e processar os sinais individu- almente, levando a soluções simples e, às vezes triviais. Como esperado, a qualidade do resultado dependerá da qualidade do sinal de entrada.

Por outro lado, se os sinais provenientes desses três sensores são amostrados ao mesmo tempo e processados adequadamente, as deficiências de cada sensor podem ser superadas e uma saída mais útel pode ser obtida. Algoritmos de tratamento de dados são utilizados para obter resultados mais sofisticados em que a saída é mais significativa do que a soma dos dados individuais. Uma solução que combina dados desses três sensores (acelerômetro 3D, giroscópio 3D e magnetômetro 3D) é chamada de solução 9 DoF (*nine degrees of freedom* – nove graus de liberdade). Aplicações que envolvem bússola, nave- gação mais sofisticas e jogos 3D são alguns exemplos que fazem uso da técnica 9 DoF e fornecem aos usuários uma experiência mais sofisticada.

É importante observar que podemos continuar esse processo no mesmo nível ou em níveis superiores e mais amplos. Por exemplo, se adicionarmos um dado de um outro sensor, como um barômetro, à técnica 9 DoF, temos uma solução 10 DoF. Neste caso, podemos usar o barômetro para detectar a altitude entre dois andares em um edifício que levam a uma navegação mais sofisticada, como é o caso de *smartphones* baseados em Windows 8.

O fato é que temos uma infinidade de dispositivos móveis com uma grande va- riedade de sensores embutidos que produzem diferentes tipos de dados sobre o mundo físico, incluindo nós mesmos, como a nossa pressão arterial. Por outro lado, temos os usu-

[1]http://en.wikipedia.org/wiki/Microelectromechanical_systems 2http://pt.wikipedia.org/wiki/Microtecnologia

ários que podem fornecer dados de participação baseados em seus contextos físico e lógico, como por exemplo, ao fazerem um *checkin* em uma rede social como Foursquare. Ou seja, o usuário de forma deliberada fornece um dado chamado de social, dando origem a um sensoriamento social (*social sensing*). Esta ampla gama de dispositivos e de dados sensoriados está representada por uma barra horizontal na parte inferior da Figura 1.

No lado esquerdo, temos dispositivos tradicionais de sensoriamento físico (senso- res) que têm capacidade de sensoriar literalmente centenas de diferentes variáveis físicas. À medida que caminhamos para a direita a complexidade de sensoriamento aumenta, como é o caso dos sensores virtuais (por exemplo, um sensor de fogo). Esse sensor pode ser construído pela combinação de outros sensores físicos como o de tempera- tura, umi- dade relativa do ar, concentração de monóxido de carbono e atividade no infravermelho. Assim, se a temperatura, a concentração de monóxido de carbono e a atividade no in- fravermelho estiverem aumentando e a umidade relativa do ar estiver diminuindo, então, dependendo do local (outro dado que pode ser obtido de um sen- sor) a probabilidade que esteja ocorrendo fogo aumenta. Sensores virtuais normalmente são obtidos pela definição de uma condição (predicado) cujas variáveis são obtidas por sensores físicos.

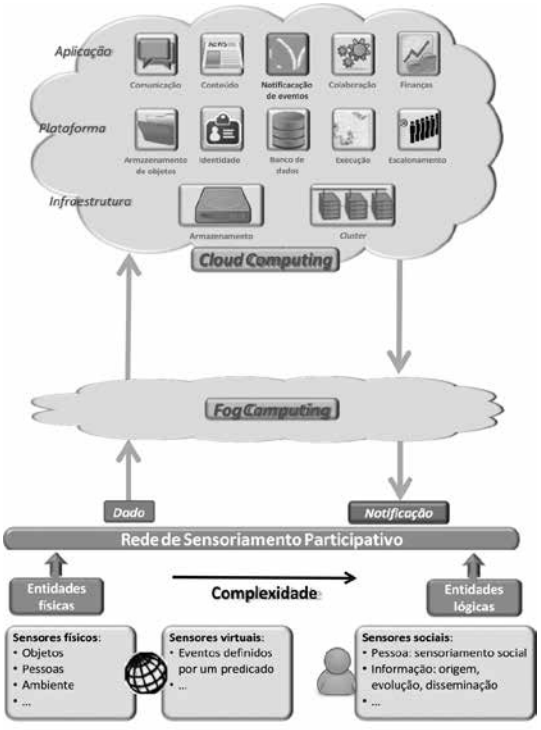


Figura 1. Rede de sensoriamento participativo

Ao caminharmos mais ainda nesse espectro, temos os sensores lógicos onde um dos exemplos mais importantes atualmente é o sensor social. Ou seja, em um extremo temos o sensoriamento de entidades físicas, normalmente associadas ao mundo físico (e.g., sensoriamento ambiental), as quais tendem a ser mais simples, passam por sensores virtuais, até o sensoriamento social, que tende a ser mais complexo em termos dos dados coletados.

Note que esses dados serão cada vez mais disponibilizados em um ambiente de computação em nuvem (*cloud computing*), utilizando uma infra-estrutura de comunicação sem fio. Na nuvem, esses dados podem ser processados e, dependendo das aplicações, notificações serem enviadas para o usuário e/ou dispositivo, como ilustrado na Figura 1.

QUESTÕES

O tratamento de dados de vários sensores heterogêneos, juntamente com o contexto no qual esses dados foram obtidos, é que criam os grandes desafios científicos para o projeto dessas redes. Certamente é onde teremos as maiores oportunidades de pesquisa conside- rando o paradigma da computação móvel, pervasiva e ubíqua, que estará cada vez mais presente nas diferentes atividades que exercemos no dia-a-dia. É esse o foco desta pro- posta de pesquisa.

O fato é que temos uma infinidade de dispositivos com uma grande variedade de sensores embutidos que produzem diferentes tipos de dados sobre o mundo físico, incluindo nós, seres humanos. Por outro lado, temos os usuários que podem fornecer dados de participação baseados em seus contextos físico e lógico, como por exemplo, ao fazerem um *checkin* em uma rede social como o Foursquare. Ou seja, o usuário de forma deliberada fornece um dado chamado de social, dando origem a um sensoriamento social.

Esta ampla gama de dados sensoriados serão provenientes de diferentes "tipos" de rede como as redes de sensores tradicionais, Intenert das Coisas, redes móveis ad hoc formadas por pessoas, veículos, objetos aéreos e aquáticos, dentre outras possibilidades.

Esses dados coletados possuem algumas propriedades interessantes. São bastante heterogênos, podem ser incorretos e incompletos, são dinâmicos (podem variar bastante ao longo do tempo) estão associados a um contexto físico (espacial e temporal) e lógico (podem depender de propriedades e perfis de pessoas). O tratamento desses dados pode ser feito perto do local onde são coletados em um ambiente chamado atualmente na lite- ratura de "*fog computing*", ou mais longe desse local em um ambiente de computação em nuvem (*cloud computing*).

Independemente de onde esses dados são processados, existem algumas questões fundamentais a serem tratadas. Dentre elas, podemos citar:

- caracterização estatística dos dados,
- fusão de dados heterogêneos,
- localização e rastreamento de entidades,
- computação sensível ao contexto,
- segurança e privacidade,
- gerenciamento de energia e
- *cloud offloading*.

A caracterização estatística dos dados revela propriedades estatísticas dos dados que permitem identificar a forma mais apropriada de usá-los. A fusão de dados heterogênos trata do processo de obter uma informação mais "valiosa" (se compararmos com os dados originais), mas considerando fontes bem diversificadas. A localização e o rastrea- mento de entidades permitem conhecer e predizer onde as entidades envolvidas em uma dada computação estão ou estarão considerando diferentes aspectos. O contexto define características que individualizam uma entidade ou um conjunto de entidades e permite o tratamento adequado do ponto de vista computacional. A segurança e a privacidade tratam de como assegurar o acesso correto, seguro e de acordo com regras previamente definidas a informações das entidades que geram dados ou que os dados direta ou indi- retamente estão relacionados. O gerenciamento de energia trata do uso eficiente de um recurso extremamente limitado, principalmente em dispositivos móveis ou que dependem de bateria. *Cloud offloading* diz respeito ao particionamento de uma computação para ser executada fora do ambiente do dispositivo seja na *fog* ou na nuvem.

O tratamento de todas essas questões, para os diferentes tipos de redes existentes individualmente (por exemplo, redes de sensores, Internet das Coisas, redes veiculares, redes móveis ad hoc, etc), como ilustrado na Figura 2, criam grandes desafios científicos que precisam ser tratados para que a academia e a indústria possam projetar soluções, serviços e aplicações que alcancem os objetivos propostos. Esses desafios se tornam ainda maiores quando temos que considerar não apenas um único tipo de rede, mas diferentes tipos já que as entidades (por exemplo, pessoas) tendem a usar mais de uma rede ao longo do tempo.

Nesse novo paradigma de computação pervasiva e ubíqua, é certamente onde o País terá as maiores oportunidades já que nós não fazemos parte do *mainstream* de países

que projeta e desenvolve a tecnologia dos dispositivos em si. No entanto, temos capaci- dade e competência para poder competir no avanço do estado da arte dessas tecnologias.

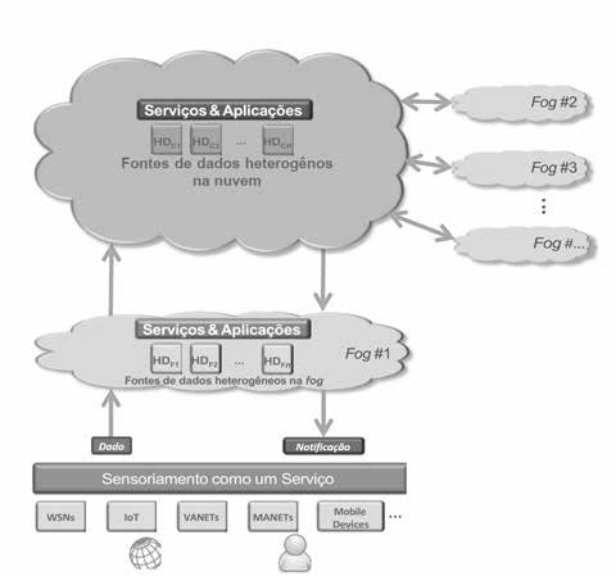


Figura 2. Sensoriamento como um serviço em diferentes tipos de redes

ASPECTOS BÁSICOS

Para tratar os desafios acima, devemos observar quatro aspectos mais básicos ainda: sen- soriamento, integração, análise e atuação. Esses aspectos estão descritos abaixo e podem ser considerados os "blocos básicos" para tratar os desafios em redes de sensoriamento participativo.

Sensoriamento

O sensoriamento (ou coleta de dados) tem por objetivo obter, de forma simples e contínua, amostras de múltiplas fontes de informação que vão desde sistemas naturais, tais como a Floresta Amazônica, até sistemas existentes em grandes cidades. Amostras de dados po- dem ser obtidas de fontes dinâmicas e heterogêneas, incluindo sensores físicos, sensores virtuais e sensores sociais, como ilustrado na Figura 1. Além do contínuo crescimento da Web e da explosão de redes sociais online, o barateamento e a modernização do hardware de sensoriamento tem levado a um crescimento sem precedentes na quantidade de fluxos de dados disponíveis em tempo real. Nesse cenário, a moni-

toração eficiente de volumes tão grandes de informação é um problema em aberto. Esta etapa trata do desenvolvimento de mecanismos eficientes para a observação do mundo físico como um repositório de in- formações sujeito a mudanças contínuas. Grandes desafios estão atrelados a esta etapa, tais como:

• Como projetar sistemas de coleta de dados que lidem, de forma eficiente, com os compromissos entre a representatividade da informação obtida e o custo, em termos de energia, espaço, latência e financeiro dos mecanismos usados para obtê- la?

• Que mecanismos devem ser usados ou desenvolvidos para coletar informação pro- veniente de fluxos de dados muito grandes, ruidosos e sujeitos a erros, levando em consideração, restrições de segurança e privacidade?

• Como permitir e encorajar usuários a compartilhar informação e assegurar a sua privacidade, de modo que dados representativos e não tendenciosos possam ser obtidos?

Integração

Na fase de integração de dados, o objetivo é projetar algoritmos e estruturas de dados que processem e combinem diferentes tipos de dados a diferentes níveis de abstração (por exemplo, texto, imagens, vídeos, gestos e ações) para extrair informação útil. Para tal objetivo seja alcançado, fazem-se necessários algoritmos que lidem com fluxos de dados massivos e que disponham de operações como agregação, filtragem e indexação em tempo (quase) real. Integração é, portanto, uma fase crítica desse processo, pois a informação constitui os alicerces sobre os quais modelos e mecanismos de atuação serão construídos.

As tarefas e problemas discutidos nesta etapa suscitam vários desafios de pesquisa, alguns dos quais são listados a seguir:

• Como integrar múltiplas fontes de dados, heterogêneas e complexas, em diferentes níveis de abstração?

• Como projetar algoritmos que sejam capazes de armazenar, agregar, filtrar e inde- xar os dados coletados de forma eficiente?

• Como avaliar a qualidade da informação derivada dos dados agregados?

• Como alcançar os três objetivos anteriores preservando a privacidade dos indiví- duos?

Análise

Outro objetivo fundamental é projetar novos modelos para descrição e predição de comportamento individual e coletivo em cenários que podem ser dinâmicos, complexos

e heterogêneos. Esses modelos devem ser capazes de receber grandes quantidades de infor- mação como entrada, informação essa proveniente de múltiplas fontes e sujeita a vários níveis de incerteza e redundância. É necessário avançar o estado-da-arte nos campos de computação natural, otimização, aprendizado de máquina e recuperação de informação, a fim de explorar e utilizar a informação proveniente de múltiplas fontes, disponível em diferentes formatos e existente em diferentes níveis de qualidade. A ideia é que esses modelos possam direcionar tomadas de decisão e atuação eficientes. Nesse contexto, os maiores desafios são:

• Como desenvolver modelos parcimoniosos, porém gerais e precisos, considerando-se a grande quantidade de características humanas e ambien- tais que precisam ser modeladas?

• Como construir modelos eficientes que forneçam respostas em tempo (quase) real? Como adaptar esses modelos para lidar com mudanças contextuais em tempo (quase) real?

• Como validar os modelos propostos, uma vez que eles podem ser alimentados com dados incompletos e muitas vezes ruidosos?

Atuação

A última etapa é o projeto e implantação de mecanismos de atuação capazes de interagir com a sociedade e com o ambiente. O principal desafio aqui é o risco de desastres. Uma decisão errada tomada por um dispositivo que controla veículos autônomos ou robôs, que aloca policiais, ambulâncias ou bombeiros, que controla o trânsito, ou guia decisões go- vernamentais, pode comprometer seriamente o bem-estar social, o ambiente e a segurança pública. Mecanismos de atuação precisam ser confiáveis, pois seus serviços podem ser requisitados a qualquer momento sob condições adversas. Esses mecanismos devem também se integrar com o ambiente e com a sociedade de forma suave, bem como serem capazes de aprender e se adaptar continuamente. A fim de desenvolver mecanismos de atuação que sejam confiáveis, resistentes e não intrusivos, deve-se responder às seguintes questõoes:

• Como construir uma vasta infra-estrutura de custo efetivo que seja capaz de pro- ver coleta de dados, computação, comunicação e armazenamento em situações adversas, preservando a segurança e privacidade?

• É possível o desenvolvimento de arcabouços de teste que permitam a quantifica- ção das limitações dos mecanismos de atuação, a fim de que sejam identificados cenários e ambientes em que tais mecanismos sejam confiáveis?

• Como isolar o impacto de diferentes mecanismos de atuação, coletar feedback, e rapidamente adaptar tais mecanismos para lidar com mudanças no comportamento de seus usuários ou do ambiente em que são utilizados?

COMENTÁRIOS FINAIS

A crescente disponibilidade de poder computacional e o surgimento de ferramentas cada vez mais expressivas possibilita a coleta de dados que quantificam de forma confiável a complexidade de redes de sensoriamento participativo. Modelos computacionais dirigi- dos a dados emergiram como ferramentas apropriadas para estudar diversos fenômenos que vão desde o aparecimento de epidemias até a propagação de informação em redes sociais. As questões tratadas neste documento têm por objetivo apresentar desafios fun- damentais para avançarmos o estado-da-arte em redes de sensoriamento participativo.

REFERÊNCIAS

Boukerche, A., Loureiro, A. A. F., Nakamura, E. F., Oliveira, H. A. B. F., Ramos, H. S., and Villas, L. A. (2014). Cloud-assisted Computing for Event-driven Mobile Services. *Mobile Networks and Applications*, 19(2):161–170.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A Survey of Mobile Phone Sensing. *IEEE Communications Magazine*, 48(9):140–150.

Silva, T. H., Melo, P. O. S. V. D., Almeida, J. M., and Loureiro, A. A. F. (2014). Large- Scale Study of City Dynamics and Urban Social Behavior Using Participatory Sensing. *IEEE Wireless Communications*, 21(1):42–51.

Wang, X., Kwon, T. T., Choi, Y., Wang, H., and Liu, J. (2013a). Cloud-assisted Adaptive Video Streaming and Social-Aware Video Prefetching for Mobile Users. *IEEE Wireless Communications Magazine*, 20(3):72–79.

Wang, X., MinChen, Kwon, T. T., Yang, L. T., and Leung, V. C. M. (2013b). AMES- Cloud: A Framework of Adaptive Mobile Video Streaming and Efficient Social Video Sharing in the Clouds. *IEEE Transactions on Multimedia*, 15(4):811–820.

SISTEMAS EDUCACIONAIS INTELIGENTES

\Carla V. M. Marques, José Otávio P. Silva, Maira Monteiro Fróes, Priscila M. V. Lima,
Claudia L. R. da Motta e Carlo E. T. Oliveira

Abstract: *Intelligent Educational Systems are computer environments where education is based and promoted by scientific principles. The intelligent education process encompasses the use of computational applications conveying metacognitive rules and interacting adaptatively with the student. Several intelligent processes collect individual and collective data from the students to compute their cognitive signature. This paper describes the specifications necessary to build intelligent educational systems. A dimensional model define measurements and requisits to build an educational system that will adapt and cater for the needs of each individual, preserving the neurodiversity of human kind.*

Resumo: *Sistemas Inteligentes educacionais são ambientes de computador onde a educação é baseada e promovida por princípios científicos. O processo de educação inteligente engloba o uso de aplicativos computacionais que transmitem regras metacognitivas, interagindo adaptativamente com o aluno. Vários processos inteligentes coletam dados individuais e coletivos dos alunos para calcular sua assinatura cognitiva. Este documento descreve as especificações necessárias para construir sistemas educacionais inteligentes. Um modelo dimensional define medidas e requisitos para construir um sistema educacional que vai se adaptar e atender às necessidades de cada indivíduo, preservando a neurodiversidade da espécie humana.*

CONCEITUAÇÃO

Sistemas educacionais inteligentes são ambientes que dominam a experiência de aprendizado do aluno na totalidade ou grande parte do processo educacional. Estes sistemas se ocupam de monitorar e coletar todas as informações recebidas do aluno individualmente e socialmente. O monitoramento deve resultar em ações imediatas que aproveitam janelas de oportunidade e criam as situações propícias de aprendizado.

Quando realizamos a análise heurística dos resultados de capturas cognitivo-linguísticas encontramos modelos refinados de ações psíquicas (operatividade observável). Diante dos desafios oferecidos pela *educação de base científica*, encontramos *clusters* de padrões decorrentes de perfis. Estes perfis são constituídos pela disposição

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE) Universidade Federal do Rio de Janeiro
Caixa Postal 15.064 – 91.501-970 – Rio de Janeiro – RJ – Brazil.

{claudiam,carlaveronica,carlo,priscila.lima,pompeuesilva}@nce.ufrj.br, froes.maira@gmail.com

de marcadores objetivos de respostas que se agrupam por semelhança. São formados pelo somatório de assinaturas cognitivas particulares dos jogadores em diferentes momentos de sua trajetória. Tais marcadores compilados como padrões, tornam-se passíveis de serem descritos e reconhecidos por intermédio de programas de *machine learning*. Marcadores são "indicadores" passíveis de serem analisados pela forma. E quando combinados, revelam também informações emergentes sobre processos cognitivo-linguísticos simultâneos e sucessivos não diretamente observáveis e que ainda são um limite para a ciência.

O modo personalizado de cada educando usar suas estratégias operativas para percorrer e decifrar as relações entre diferentes redes epistêmicas implícitas no jogo, viabiliza por meio da inteligência artificial, o entendimento da transitividade do pensamento em tempo real. Desta forma o processo educacional reage adaptativa e evolutivamente de acordo com as atitudes do educando quando este cria soluções previsíveis em conformidade com um gradiente matematizado de diferentes níveis e tipos. O encadeamento de procedimentos psicológicos e aplicações de esquemas formam um entrelaçamento de relações de regras generativas no pensamento e que se realizam através da migrações de invariantes transpostos entre diferentes áreas do conhecimento.

Desta forma homens e computadores interelacionam-se, elevando-se mutuamente. Os computadores compreendem, interpretam, se adaptam e evoluem para responder em tempo real às intervenções e estímulos que recebe. Da mesma forma o cérebro humano desafiado, precisa se tornar incremental para continuar interagindo e criar desafios em novos patamares de complexidade para a máquina.

Descrever e conhecer os padrões mais sofisticados da mente humana, por conta de sua natureza dinâmica (Inhelder, 1996), permite o homem ser transformado pela máquina e dialeticamente modificá-la criando uma interação infinita: cérebro-mente humana e hardware- software computacional, isto significa identificar um emaranhamento quântico homem-máquina.

Com posse desses resultados, os sistemas educacionais inteligentes tornam-se máquinas aprendentes que paramediam interações autônomas com objetos reais de conhecimento e quando colaborativos, intensificam as interelações entre educandos. A finalidade é provocar saltos cognitivos nos participantes. A paramediação interativa precisa ser dinâmica, buscando provocar a tomada de consciência das regras generativas implícitas nos games inteligentes. Isto significa tornar-se capaz de reaplicar operações lógico-gramaticais na forma de um grupo de estruturas canônicas constituídas por regras universais da cognição. Este processo torna o ser humano capaz de realizar a transitividade do pensamento utilizando padrões estruturais e algorítmicos entre os sistemas aplicados às diferentes linguagens do conhecimento. Por isso,

um dos atributos mais relevantes dos games inteligentes é a interação evolutiva e adaptativa no desenrolar dos seus desafios (Marques, 2009).

A educação convencional consiste em aumentar o nível de informação dos conteúdos didáticos do educando. Já nos sistemas educacionais inteligentes, o educando enfrenta desafios calculados para interferir no processo de aceleração da cognição. As informações são capturadas de modo a possibilitar a visualização das dimensões da cognição, armazenando dados pertinentes e compararando-os com as teorias da mente pré-existentes (Marques, 2009).

Cada processo educacional inteligente é único em sua especificidade e arquitetura de construção. A perspectiva neuro-pedagógica respalda o processo de ensino-aprendizagem para uma abordagem lúdica e coerente com conteúdos didáticos interativos. O design metacognitivo é o principal diferenciador desta engenharia educacional, proporcionando oportunidades de disparar informações sobre os esquemas e representações mentais de alta complexidade (Marques, 2009, Inhelder, 1996).

Esta engenharia requer a construção de instrumentos de medida específicos e modelados para funcionar durante a execução dos games. São crivos matematizados que após tratamento dos dados que coletam possibilitam a identificação de um índice dinâmico qualitativo da metacognição. Mas não há fórmulas pré-concebidas de modelos para a criação de processos educacionais inteligentes, mas uma série de etapas sucessivas e simultâneas de tarefas que incluem várias áreas de conhecimento (Marques, 2009).

A proposta dos sistemas educacionais inteligentes é realizar em tempo real, duas fases: análise da assinatura cognitiva do educando e paramediação personalizada adaptativa e evolutiva para o salto cognitivo. A análise da assinatura cognitiva revela a natureza e funcionamento psíquico. Com essas informações colhidas durante o aprendizado podemos determinar que tipo de mediação deve ser utilizada para que o educando construa novos processos mentais para realizar um salto cognitivo (Marques, 2009, Puchkin, 1969).

A maneira mais prática de se criar um Sistema Educacional Inteligente é através da gamificação do processo educacional. Para isso, um processo de desenvolvimento específico é aplicado para construir um game que não só seja capaz de ensinar uma determinada competência mas também rastreie a interação segundo um modelo estabelecido. O processo visa garantir que o aprendizado não seja superficial e que o modelo do game seja capaz de se adaptar monitorando a evolução do aluno.

PROCESSO DE CRIAÇÃO DE SISTEMAS EDUCACIONAIS INTELIGENTES

A construção de Sistemas Educacionais Inteligentes requer uma engenharia de desenvolvimento baseada em um modelo de pesquisa científica. O tema a ser aprendido é baseado em um sólido referencial teórico para que resulte em um modelo dimensional relevante. Com o espaço dimensional definido, o processo educacional pode ser inventado, levando em conta os preceitos determinados pelo estudo das bases científicas. Neste processo criativo, a arte inicial do processo educacional é pareada com os requisitos do modelo cognitivo, definindo os episódios de aprendizado. A pedagogia fica a cargo de um estudo das possíveis interações do educando e o significado das respostas dentro do conteúdo ensinável. Todas estas informações são reunidas para dar corpo final ao processo educacional, agregando as regras que serão estabelecidas para o educando e para o engenho que representa o modelo matemático da teoria.

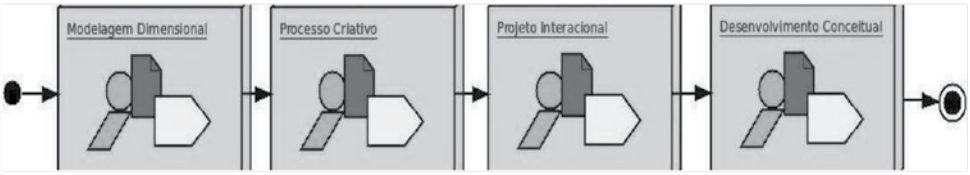


Figura 1. Processo de Criação do Sistema

ESQUEMA DO PROCESSO DE CONSTRUÇÃO

Os documentos são gerados em diversas atividades envolvendo uma equipe interdisciplinar. As atividades iniciais são formalizações das teorias embasantes feitas por cientistas da educação e especialistas do tema abordado pelo processo educativo em particular. O especialista do tema, que pode ser um professor de sala de aula, leva os conceitos modelados para artistas traduzirem estes conceitos em linguagem, música e imagem. Profissionais de ensino estudam as potenciais respostas que se esperam dos educandos e ponderam a relevância destas respostas no escopo do conhecimento que está sendo estudado. O processo educacional poderá ser finalizado por uma equipe de designers, engenheiros e desenvolvedores.

O MODELO DIMENSIONAL

Todo game é capaz de trabalhar a cognição humana em diversas dimensões. Estas dimensões devem ser consideradas dentro da visão cognitiva que se quer mensurar e intervir. O espaço dimensional deve ser projetado de modo a conter todas as dimensões relevantes e ao mesmo tempo deve reduzir ao máximo estas dimensões em uma simplificação de engenharia. O modelo dimensional é o espaço onde se desenvolve o

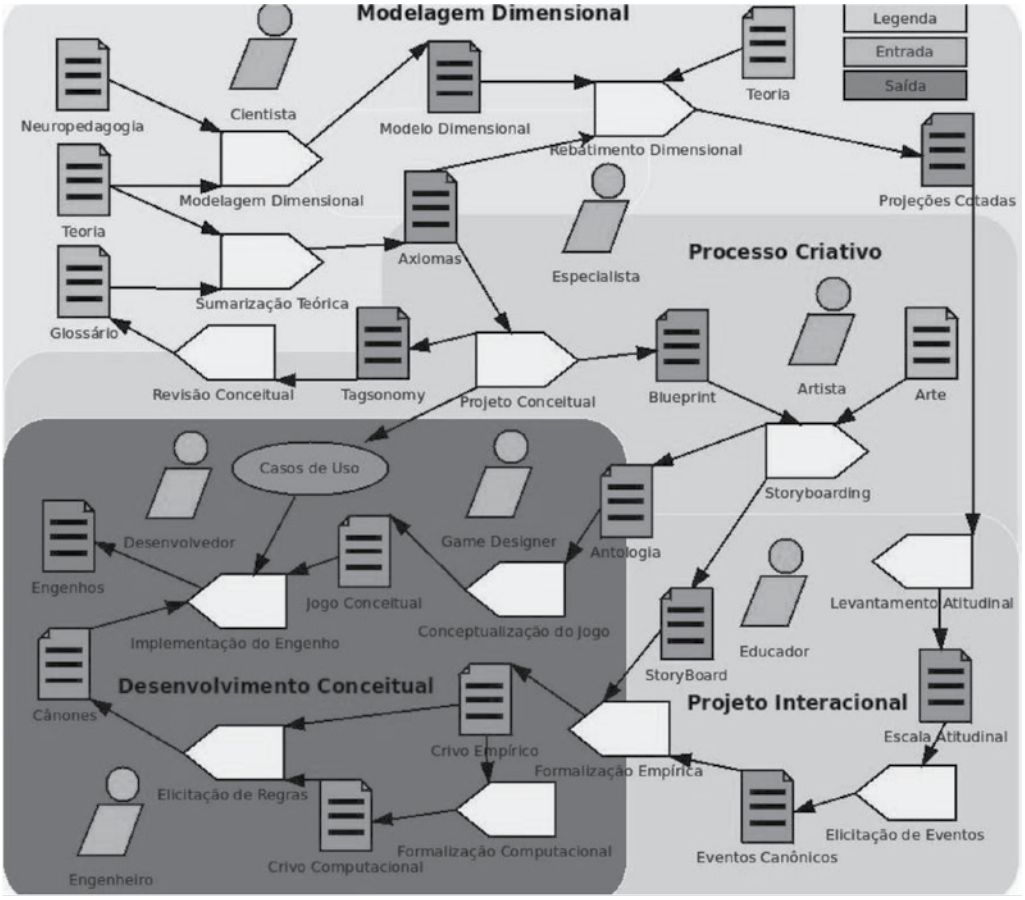


Figura 2. Processo Completo da Construção de um Sistema Educacional

itinerário do sujeito aprendente. Este espaço é coordenado pelos eixos ortogonais representando cada dimensão nas magnitudes previstas. Cada ponto no espaço representa a competência do indivíduo naquele momento. Os deslocamentos no espaço representa transições entre competências ou estados competentes.

Magnitudes/dimensões	Descontínuo	Contínuo
Números	Inteiros	Reais
Operações	Multiplicação	Divisão

Tabela 1. Modelo Dimensional do Domínio da Matemática

Modelo dimensional deve ter domínio, ortogonalidade e grandeza. No exemplo se faz o recorte do domínio da matemática em duas dimensões ortogonais entre si, o contínuo e o descontínuo. As grandezas destas dimensões são competências a serem adquiridas na matemática. No exemplo é presente o recorte com apenas as duas grandezas números e operações.

Os desafios da modelagem dimensional é o extenso trabalho de criação necessário para cobrir a educação mínima e a adequação destes modelos para a criação de contrapartidas computacionais. O tamanho da tarefa propicia que ela seja abordada por diversos laboratórios e pesquisadores de áreas interdisciplinares. É uma tarefa que nunca será concluída, pois a medida que o conhecimento avança, novos modelos ou atualizações em modelos existentes irão surgir. A aderência do modelo à uma implementação computacional efetiva requer um estudo aprofundado de vertentes filosóficas, científicas e de engenharia. O modelo deve ser capaz de atender às expectativas de um Sistema Educacional Inteligente, possibilitando que o estudante seja compreendido em suas necessidades e paramediado adequadamente. A teoria e o processo que garanta isso é algo que precisa ainda ser pesquisado. Isto abrirá novas frentes de pesquisa dentro da área de computação aplicada e também na criação de uma especialidade da engenharia de software que tornasse uma engenharia educacional computadorizada.

PROCESSO CRIATIVO

O modelo dimensional combina teoria da educação com a engenharia para formar a base de construção de um modelo conceitual. O modelo conceitual é um conjunto de especificações sob diversas visões delimitando um produto educacional contextualizado no modelo dimensional construído.

O modelo conceitual requer que se enumere um conjunto de axiomas definindo os aspectos do produto educacional em alto nível. Os principais axiomas se originam da área teórica subjacente à carga útil do produto educacional. Outras axiomas são de áreas mediacionais que vão garantir que a carga útil atinja o seu propósito. Os requisitos resultantes deste detalhamento definem as interações que o educando faz com o processo educacional e as reações que o processo pode apresentar. Um blueprint pode ser construído com um simples relato textual em parágrafos ou itens de tudo que se pretende construir no aprendizado. O modelo mais completo e formal do blueprint é o um modelo de casos de uso. O modelo feito em casos de uso podem ter diversos níveis de detalhamento segundo as necessidades de refinamento.

A antologia é uma coleção de histórias minimalistas ou microepisódios que descrevem cada um dos requisitos da planta baixa, relacionadas ao desenrolar dos episódios constituintes do enredo do processo educacional. O storyboard é uma amostra destas histórias contextualizada com a apresentação do ensinamento ao educando nor-

malmente em uma forma de rascunhos gráficos. As micro histórias antológicas são selecionadas a partir de diversas histórias candidatas usando como critério a taxa de relevância que elas projetam nos eixos dimensionais. As histórias são relevantes e portanto antológicas na medida em que revelam estados e processos cognitivos pertinentes ao espaço de investigação e intervenção do processo educacional. Numa antologia os microepisódios são classificados usando uma tipologia folksonômica, ou seja, palavras-chave são atribuídas a cada um deles, e pareados com os axiomas de onde se originam.

PROJETO INTERACIONAL

A construção de qualquer conhecimento humano é feita através de um conjunto de atitudes inconscientes e conscientes em um ciclo cognitivo-volitivo. Esta mecânica é conhecida como processo das microgêneses cognitivas. Estas atitudes estão dispersas ao longo de um escopo de funções cerebrais cognitivas apresentando diversos níveis de observabilidade. Estas atitudes pertencem a um módulo procedimental, que é uma sequência de operações requeridas para executar um ato cognitivo-volitivo intencional. A escala atitudinal relata conjuntos de atitudes pertinentes a cada módulo procedimental e as magnitudes que cada uma delas representa dentro do procedimento. Estas magnitudes representam a relevância de cada atitude para o reconhecimento da natureza do procedimento. A computação ponderada destas atitudes permite discriminar um conjunto de eventos que disparam a marcação de um registro em um ou mais eixos dimensionais.

Todas as interações observáveis do educando com o processo educacional são oportunidades de se registrar a sua assinatura cognitiva, a maneira como ele adquire conhecimento a partir do aprendizado. No entanto, o registro indiscriminado de todas as interações pode gerar uma massa de dados muito grande e de difícil interpretação. No projeto da planta baixa os requisitos referenciam os axiomas teóricos e determinam junto com o quadro de projeções a relevância de cada observação. O repertório de eventos canônicos é construído elencando eventos do educando ou ativados pelo processo que representam oportunidades de observação e intervenção. A coleção de eventos canônicos é a mínima coleção de eventos capaz de discriminar todos os módulos procedimentais requeridos na investigação e intervenção projetadas para um determinado processo educacional.

Em um sistema dimensional que lida com fatores humanos, nem sempre as dimensões são contínuas ou isotrópicas. As magnitudes nestas escalas muitas vezes são pontuadas por marcadores abstratos descritos por uma situação reconhecível e muitas vezes complexa. No modelo dimensional de Guilford (figura 3) vemos a dimensão de produto pontuada por magnitudes discretas como unidade, sistema, relação, transfor-

mação e implicação. O mapeamento axiomático normalmente já descreve um conjunto de marcadores que podem ser atribuídos a cada uma das dimensões pertinentes (projetadas). O crivo empírico deve relacionar conjuntos de eventos canônicos a cada uma das marcações arbitradas no modelo dimensional. O crivo empírico deve conter então a relação da escala atitudinal com a marcação dos eventos canônicos nas respectivas dimensões pertinentes. O crivo empírico é o filtro que discrimina os diversos estados cognitivos observáveis descritos nos referenciais teóricos.

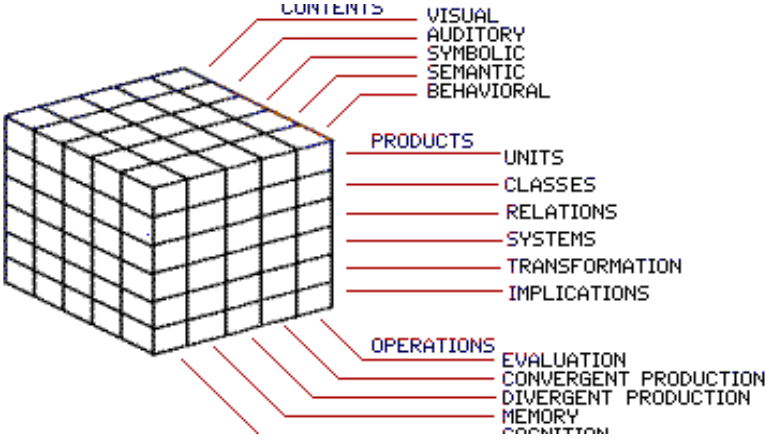


Figura 3. Teoria de Guilford sobre a Estrutura do Intelecto

Marcadores são descrições que podem definir pontos em uma ou mais dimensões. Alguns marcadores podem ser referenciados a uma dimensão e incorporar a escala atitudinal definindo pontos discretos que serão valorados no crivo empírico. Os marcadores podem ser conceitos abstratos e serem observáveis a longo prazo ou apenas por uma observação objetiva. Neste caso eles dificilmente podem ser transladados para o crivo computacional, mas servem como sistema de calibração do processo educacional quando aplicado em um ambiente controlado.

Com o avanço da neurociência e computação, novas possibilidades surgem para um tratamento mais refinado e introspectivo de todos os dados que emanam dos processos educacionais. O crivo computacional deve ser baseado na hipótese de que diversos estados e atitudes não observáveis podem ser inferidos em processos de mineração que vão além dos dados coletados. O crivo computacional avançado normalmente inclui um modelo matematizado da antologia baseado no mapeamento axiomático. Este modelo é norteado por estratagemas conceituais do raciocínio ou gambitos mentais. Gambitos mentais são estruturas-processo que condensam a habilidade de raciocinar do cérebro pensante.

O principal conceito que norteia a construção e operação de sistemas educacionais inteligentes é a transitividade metacognitiva. Este conceito define a existência de uma cognição de alto nível que resume todos os conhecimentos e capacidades em metaregras universais. O cânone emissário representa o conjunto de regras e metaregras que o aprendizado consegue eliciar no educando. Estas regras devem estar embutidas na inteligência cibernética do sistema educacional e no seu design metacognitivo para que desperte as regras correspondentes na cognição do educando. O cânone tributário ou afluente é construído quando o propósito da educação é de prótese cognitiva. O sistema educacional deve ser criado para que seja o excipiente de um conjunto de regras ou metarregras que devam ser implantadas na cognição do educando.

O cânone emissário é um conjunto de regra que tentam mapear aquelas que o educando já deve saber. O esforço de construção deste cânone está focado na formalização das regras que o aprendizado pode suscitar no educando. Estas regras devem ser enunciadas de modo que um algoritmo computacional detecte o uso delas pelo educando (tabela).

O cânone tributário é feito com as regras determinadas pela teoria que embasa o processo educacional. O crivo computacional discrimina quando o educando viola alguma regra que está definida na teoria. O engenho então deve estar programado para que o educando possa aprender a regra. O processo educacional passa a oferecer desafios ligados à descoberta da regra desejada. Uma outra opção é o uso de um processo chamado elaboração dirigida. Na elaboração dirigida, as perguntas ou desafios procuram instigar o educando a explorar uma visão do problema que ainda não foi coberta por ele.

DESENVOLVIMENTO CONCEITUAL

O desenvolvimento conceitual é feito a partir de uma descrição abrangente de alto nível de toda a mecânica e dinâmica do processo educacional. Um conjunto de invariantes são estabelecidos para representar as metaregras embutidas no processo e a partir destes invariantes é levantado o espaço de divergência oferecido para o educando.

Sistemas educacionais inteligentes são construídos dentro de uma perspectiva científica. O método científico permite a construção do sistema usando modelos matemáticos. Estes modelos matemáticos podem ser incorporados tanto no cliente como no servidor através de engenhos computacionais. Esta modelagem é o principal diferencial entre um modelo educacional tradicional e um inteligente. A pedagogia da educação convencional leva em conta apenas a transmissão de um conteúdo e sua assimilação pelo educando. O modelo e o engenho do sistema educacional inteligente incorporam conhecimentos científicos sobre o comportamento do educando e

principalmente sobre suas competências e habilidades mentais. O modelo matemático computacional descreve o comportamento esperado do educando segundo uma dada teoria e representa o processo educacional segundo este conceito hipotético. O engenho computacional é um sistema de software que incorpora na pedagogia todo o modelo matemático concebido pela teoria. O engenho se encarrega de acompanhar as interações do educando e validar segundo o modelo programado. Durante todo o processo de aprendizagem o engenho reúne os dados de todos os educandos no banco de dados e aplica o modelo definido para aspecto cognitivo estudado. O engenho faz inferências no conjunto de dados e levanta o grau de validade do modelo estudado. A parcela validada do modelo interage com o processo de aprendizagem, configurando otimizações no engenho computacional. Registros de ações do educando que forem irrelevantes podem ser suprimidos. Ações programáticas previstas pelo modelo validado podem ser ativadas para realizar novas interações com o educando.

O modelo matemático de um processo educacional inteligente pode ser obtido a partir da matematização dos axiomas teóricos ou a partir de um forma mais refinada advinda do crivo computacional. O modelo matemático é uma representação formalizada das teorias que embasam o processo educacional juntamente com os aspectos neuropedagógicos usados para garantir que as competências, habilidades e conhecimentos sejam assimilado pelo educando. A teoria já tem a sua primeira representação no mapeamento axiomático. Os axiomas devem ser acompanhados por uma formalização que permita a sua incorporação em um engenho computacional. Os conceitos teóricos são mapeados em blocos funcionais do modelo que são responsáveis pelo acompanhamento do processo neuropedagógico educacional segundo os preceitos teóricos. Os aspectos neuropedagógicos são aqueles diretamente ligados à monitoração e controle da aprendizagem, em um modelo de metacognição artificial. O modelo neuropedagógico reúne requisitos advindos dos crivos e do cânone elaborativo para definir uma representação matemática do processo educacional e do educando.

O engenho computacional tem o propósito de implementar no processo educacional o modelo matemático e vários outros aspectos definidos pelos documentos neuropedagógicos. O engenho computacional tem a sua origem no modelo matemático e incorpora outros requisitos refinados advindos de documentos neuropedagógicos mais detalhados. O engenho é um sistema que envolve diversas partes que refletem a complexidade do processo educacional implementado(figura 4). Cada processo recebe um engenho cliente unitário que implementa os cânones de regras definidas na especificação do sistema educacional inteligente. O cânone tributário controla as respostas que o cliente dá quando os limiares prescritos pelo crivo computacional são ativados. O cânone emissário avalia as interações do educando definidas pelos eventos canônicos e envia para o engenho de coleta através do engenho conectivo. O engenho conectivo gerencia processos educacionais coletivos e gerencia o crivo

computacional coletivo através de um módulo reator. O engenho de coleta registra um documento para cada educando contendo todos os dados coletados em todos os processos educacionais que ele participou. O módulo de acesso calcula as assinaturas cognitivas dos educandos definidas no modelo matemático e provê acesso aos prognósticos neuropedagógicos para o engenho de análise e especialista. O engenho de análise executa a parte avançada do modelo matemático que incorpora inteligência computacional e outras análises que envolvem big data e não são feitas em tempo real. O engenho especialista é acoplado a um aplicativo que educador usam para interagir com os dados processados pelo sistema. Ele contem um filtro que determina as visões que o educador deseja observar de cada educando tanto em tempo real como nas assinaturas temporais. O módulo interventor tem uma versão especial do modelo matematizado que permite decisões do educador para cada educando ou para um grupo de assinaturas cognitiva. Essas intervenções são roteadas para o módulo adaptativo do Engenho unitário, afetando os processos educacionais determinados.

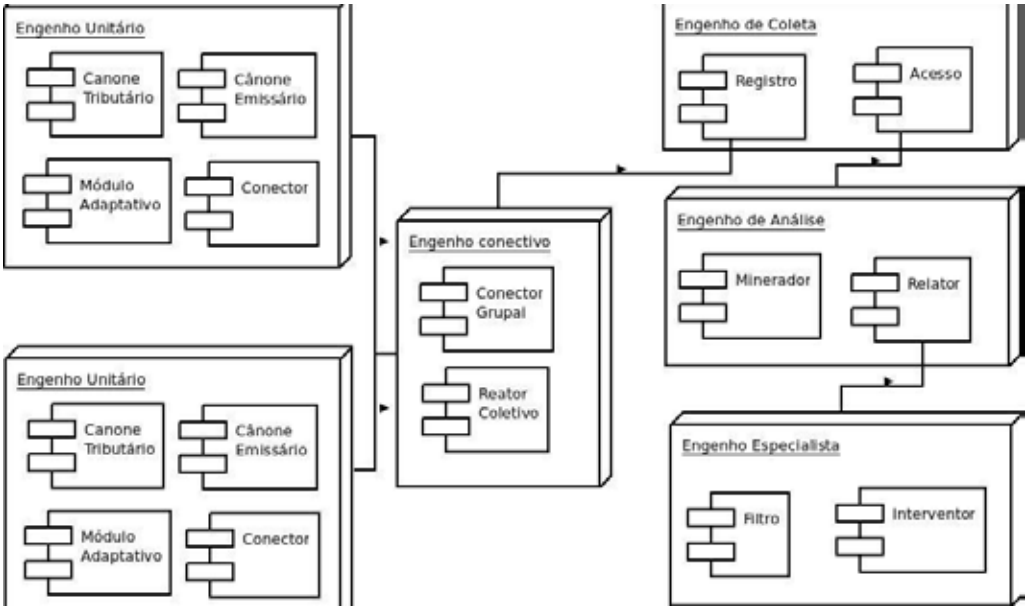


Figura 4. Componentes de um Engenho para Sistemas Educacionais Inteligentes

ANÁLISE E PARAMEDIAÇÃO

A coleta de informações do aluno proporciona uma visão estratégica da totalidade do alunado que permite a sintonização social e o traçado de grandes rumos. O núcleo do sistema de educação é um modelo da mente aprendente do aluno sinto-

nizado por aprendizagem de máquina. e um conjunto de heurísticas que adaptam o ensino ao momento e estilo de aprendizagem do aluno. O ápice de um sistema educacional inteligente é um processo gamificado onde o total do aprendizado se dá em um conjunto de situações problemas que fazem parte do enredo de um game. O game incorpora a simulação mais apropriada da situação que exige a competência e as habilidades que um aluno ou grupo tem de adquirir. Um sistema de ensino assim não precisa de uma avaliação em separado, mas a a avaliação é feita de maneira contínua, ajustando progressivamente os aspectos do problema até que ele possa ser completamente dominado pelo aluno. A avaliação é então parte integrante do aprendizado e deixa de ser a experiência frustrante de passar ou falhar. Os dados colhidas nesta avaliação servem para aperfeiçoar a inteligência do sistema que vai se tornando mais eficiente na adaptação do modelo personalizado de aluno.

Informações coletadas em um sistema educacional inteligente são oriundas de todo o tipo de interação do educando com o processo educacional. Os sistemas clientes devem enviar um conjunto de dados recolhidos para o servidor, segundo os crivos que estão programados nele. Inicialmente uma interface padronizada de coleta é definida para todo e qualquer processo educacional (tabela 4).

Coluna/quesito	Descrição	Relevância
Marcador	Objeto referenciado na interação	Seleção de um ou mais interadores
Posição	Localização terminal do objeto	Decisão do objetivo a ser alcançado
Ação	Modelo de interação usado	Regra generativa aplicada
Pontuação	Índice de sucesso calculado pelo crivo	Correlação da regra usada com a teórica
Tempo	Tempo usado para fazer a interação	Assinatura temporal do processo mental
Resultado	Estado cognitivo registrado pelo crivo	Assinatura espacial do processo mental

Tabela 2. Interface Padronizada de Coleta

Estes dados são o mínimo requerido para se obter uma assinatura cognitiva pertinente. No entanto, processos educacionais mais complexos podem requerer uma interface de coleta mais extensa que deve ser definida no crivo computacional e incorporada e analisada pelo sistema central. Na interface mínima fica definida a série temporal de ações que definem as resposta dado pelo educando ao processo educacional corrente. Nesta série pode-se analisar o ganho do aluno ao longo de todo processo, definindo a receptibilidade que o processo alcançou. Em um resultado amostrado, os crivos computacionais avaliam tudo o que foi registrado e comparam com prognósticos e outros resultados classificados. Esta comparação permite selecionar mudanças

no processo que beneficiem o aprendizado deste estudante. A oferta deste novo aprendizado pode ter sido recomendada pelo embasamento teórico ou reconhecida por similaridade a outro estudante que obteve melhor resultado em um outro contexto.

Como exemplo podemos demonstrar o problema de manobragem de trens adaptado de um sugerido por Bärbel (Inhelder 1996). É um problema de lógica para estudar como uma pessoa usa seus recursos para resolver uma situação que só tem uma resposta final mas com inúmeras maneiras de resolver. Na figura vemos um trem que precisa ser invertido e uma junção que pode ser chaveada para que o trem possa seguir pela seção norte ou sul. O estudante pode manobrar o trem para frente e para trás, desengatando ou engatando porções do comboio (Figura 5).

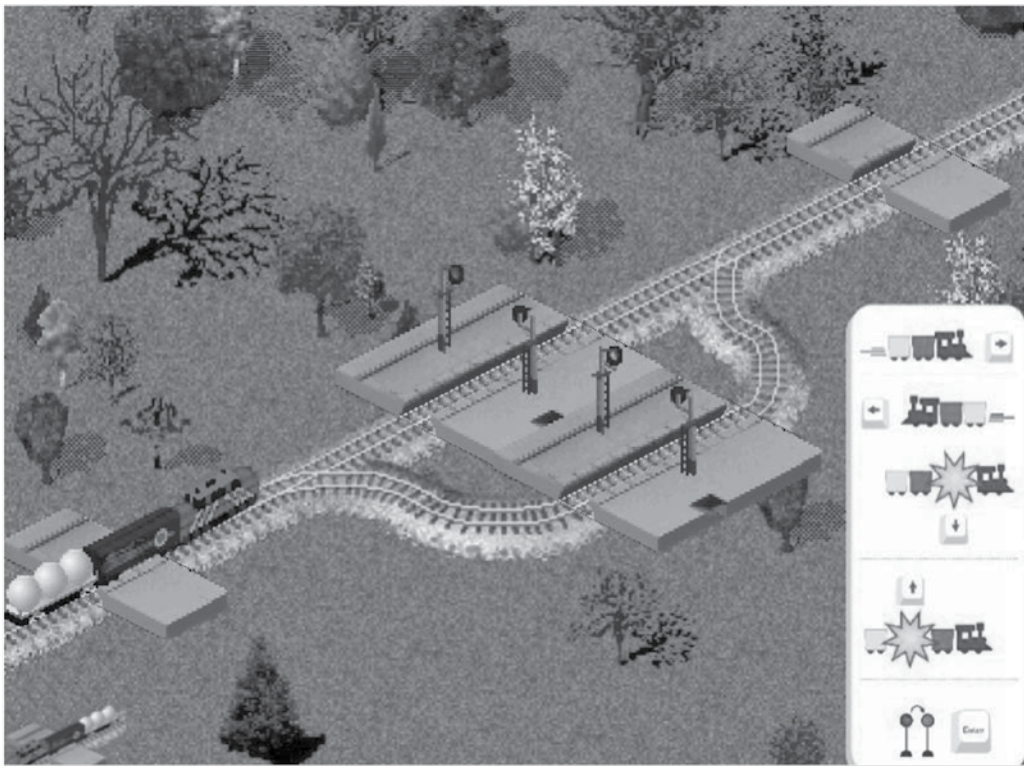


Figura 5. Problema da Manobragem de Trem

A tabela é uma amostragem do resultado deu uma sessão onde se resolve o problema. O marcador se refere à combinação do comboio onde loco é a locomotiva, good é o vagão de mercadorias e ceme é o de transporte de cimento. A posição é o segmento de trilho para onde a locomotiva se deslocou. A ação é dada por o, origem;

b, voltando; f, avançando; u, desengate acima; d, desengate abaixo e c, chaveia a junção. No resultado temos apenas dois resultados aparentes o normal e o sucesso. A pontuação é dada por um crivo que avalia a qualidade das decisões tomadas e o avanço em direção da solução. Este processo não incorpora inicialmente nenhuma paramediação, ou seja, ele não fornece nenhuma intervenção baseada no acompanhamento do estudante.

Marcador	Posição	A	Pontos	Tempo	Resultado
ceme:good:loco	FRWEST	o	0	0.0159997940063	NORMAL
good:loco	FRWEST	d	1.86666666667	6.86099982262	NORMAL
ceme:good:loco	WEST	f	2.49714285714	15.6749999523	NORMAL
loco	WEST	u	2.66666666667	18.8499999046	NORMAL
ceme:good:loco	WEST	b	3.11794871795	28.3699998856	NORMAL
ceme:good:loco	SOUTH	f	3.17142857143	32.4289999008	NORMAL
loco	SOUTH	u	3.21176470588	37.1599998474	NORMAL
loco:ceme:good	SOUTH	f	3.75652173913	54.0569999218	NORMAL
loco:ceme:good	EAST	f	3.80598290598	61.7329998016	NORMAL
loco:ceme	NORTH	f	4.21287878788	73.8559999466	NORMAL
loco	WEST	b	4.43962703963	78.4519999027	NORMAL
loco:good	WEST	c	4.25044722719	85.3039999008	NORMAL
loco:good:ceme	NORTH	f	4.12305194805	86.7979998589	NORMAL
loco:good:ceme	WEST	b	4.0784057971	89.0829999447	NORMAL
F_I_N_I	F_I_N_I		1	2014-09-14 20:04:55.742824	SUCCESS

Tabela 3. Amostragem de Coleta de Dados do Problema Trem

Os dados coletados fornecem informações sobre o aprendizado que o estudante está fazendo em tempo real. Os crivos computacionais podem avaliar o quanto o estudante está avançando ou quais são os obstáculos que ele está enfrentando. No gráfico 1 podemos avaliar o comportamento do aluno diante do problema. Ao observar

o movimento entre os tempos 15 e 60 percebemos que houve uma repetição padronal do movimento de vai e vem sem que a configuração do marcador se altere muito. Isto significa que o estudante está em dúvida não entendeu ainda o que deve ser feito para solucionar o problema. A partir do tempo 60 temos uma evolução significativa do marcador e sua pontuação tem uma nova derivada. Isto sinaliza que o educando achou a solução e está agora aprendendo mais rápido.

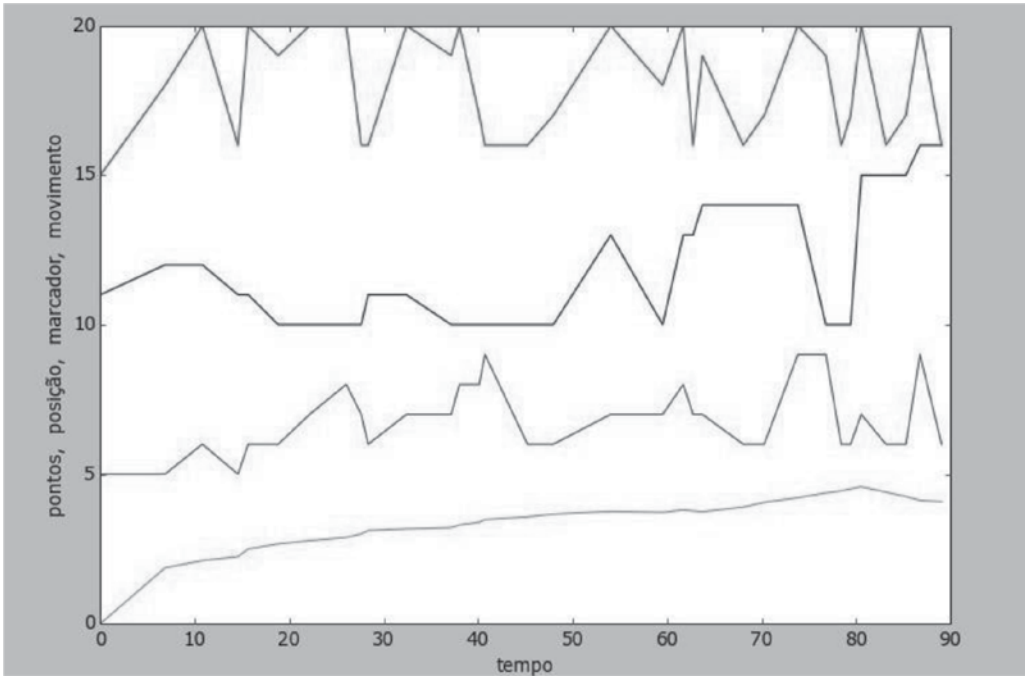


Gráfico 1. Evolução da Manobragem do Trem

Um sistema de paramediação feito com uma combinação de perfis prognósticos e coletados pode intervir no sistema educacional inteligente. Enquanto a exploração do problema estiver dentro de parâmetros populacionais aceitáveis não deve haver nenhuma intervenção. O sistema inteligente usa diversas heurísticas e recursos da inteligência computacional para extrair conclusões dos dados coletados. A medida que se constrói modelos mais precisos do comportamento, maior vai ser afetividade da resposta do sistema. Modelos matemáticos que levem em consideração maior número de parâmetros e que sejam auto adaptativos à evolução do dados e perfis populacionais serão mais eficientes no lidar com o educando. Outra face do problema é a capacidade deste sistema se reportar aos educadores dando informações precisas

sobre as condições de cada aluno ou grupo. O sistema deve ser capaz de identificar os problemas e encontrar soluções, indicando a um professor ou a um aluno que possa ajudar na evolução do educando em dificuldades.

Ao se constatar que o estudante precisa de auxílio o sistema pode intervir mudando a dificuldade do problema ou ampliando a oferta exploratória. No problema do trem um outro trem em uma linha paralela pode começar a manobrar em um problema diferente mas que sugira aspectos de solução que não foram explorados. Caso o estudante evolua muito rápido para a solução, um outro trem pode se aproximar e engatar no comboio aumentado a dificuldade do problema. Caso este problema se mostre insolúvel para este estudante, ele passará a não ser indicado para alunos que tenham o perfil semelhante. O problema deverá ser apresentado de outra forma mas com uma lógica similar. Por exemplo, em vez uma forma gráfica o problema pode ser escrito num texto ou narrado em áudio.

CONCLUSÃO

Criar uma educação inteligente é um grande desafio da computação. A modelagem de um cérebro aprendente é a solução para que se tenha o total aproveitamento do processo de aprendizagem que se dê em uma situação problema real. Uma educação que gere indivíduos que saibam operacionalizar o seu conhecimento e não apenas memorizar ou imitar. Este é o verdadeiro papel da computação na educação, criando mecanismos adaptáveis e evolutivos que possam preservar a neurodiversidade da espécie humana, levando cada indivíduo ao máximo de sua potencialidade.

Um sistema inteligente aprende o processo aprendente de cada indivíduo e ensina de maneira que cada lição seja uma coisa nova e desafiante dentro das expectativas e qualidades que cada ser individual tem. Este modelo também valoriza o professor que deixa de ser a figura que premia ou castiga quem aprendeu uma determinada matéria e passa a ser o que encaminha o indivíduo na realização plena do seu ser. O modelo computacional cria os problemas e o professor oferece pessoalmente ou através do sistema a orientação para o indivíduo avançar na solução.

REFERÊNCIAS

Broch, José Carlos. O Conceito de Affordance Como Estratégia Generativa No Design de Produtos Orientados para Versatilidade. UFRGS, 2010.

Dehaene, S. Os Neurônios da Leitura. Penso: Porto Alegre, 2012. Guilford, J.P.. The Nature of Human Intelligence, 1967.

Inhelder, Bärbel [et al.]. O Desenrolar das descobertas da criança: pesquisa acerca das microgêneses cognitivas. Trad. Eunice Gruman. Porto Alegre: Artes Médicas, 1996.

Kienitz M.L.Modelo fractal das microgêneses cognitivas: uma metodologia para a mediação metacognitiva em jogos computacionais, SBIE 2012.

Maddux, C.D. & Gibson, D. Research Highlights in Technology and Teacher Education 2012. SITE. Retrieved January 13, 2014 from http://www.editlib.org/p/41222. 2012.

Marques, Carla Verônica; Oliveira, Carlo E.T. de; Motta, Cláudia (Org.). [et al.]. A Revolução Cognitiva: um estudo sobre a teoria de Franco Lo Presti Seminério. Instituto de Matemática. Núcleo de Computação Eletrônica. Relatório Técnico 04/09. Rio de Janeiro. 2009.

Puchkin, V.N. Heurística A Ciencia do Pensamento Criador. Rio de Janeiro: Zahar Editores, 1969.

Relatório Técnico UFRJ/PPGI, A Máquina da Metacognição, Neuropedagogia II, 2010.

Shimamura A. e Janet Metcalfe, Metacognition: Knowing about Knowing. Massachusset Institute of Tecnology Cambridge, 1992.

PARTICIPAÇÃO POPULAR E TECNOLOGIAS: EXPERIÊNCIAS E DESAFIOS

Cristiano Maciel[1], Claudia Cappelli[2,3], Cleyton Slaviero[4]

Abstract. *The development of environments that promote citizens' electronic participation (e-participation) in governmental issues is a challenge that requires robust and broad architectures and methodologies, besides an effective infrastructure that fosters and supports citizens' participation. Another challenge is an education model that leads citizens to participate in public life, with transparent information and technology-meadiated engagement. Facing those issues, we herein discuss and deepen researches and practices on that subject, within the Great Challenges posed by the Brazilian Computer Society.*

Resumo. *O desenvolvimento de ambientes que promovam a Participação eletrônica (e-Participação) dos cidadãos em questões governamentais é um desafio que requer arquiteturas e metodologias robustas e abrangentes, além da construção de uma infraestrutura efetiva que estimule e suporte a participação do cidadão. Também, o incentivo a uma Educação que conduza os cidadãos a participarem da vida pública, com transparência nas informações e engajamento mediado pelas tecnologias, é um desafio. Frente a estas questões, propõe-se a discussão e aprofundamento das pesquisas e práticas deste tema como parte dos Grandes Desafios da Sociedade Brasileira de Computação.*

INTRODUÇÃO

O uso das Tecnologias da Informação e da Comunicação (TIC), em especial da Internet, possibilita ao cidadão a obtenção de serviços, o acesso a informações para geração de conhecimento e a consequente participação junto às questões governamentais, tornando real a democracia e auxiliando ambos na tomada de decisões (Maciel, 2008). Todavia, para que de fato haja participação dos cidadãos, estes devem conseguir articular um discurso, elaborar propostas, confrontá-las e indicar suas escolhas, através de meios de difusão pública. A democracia eletrônica (e-Democracia) pode viabilizar tal articulação, promovendo a discussão em torno de um processo, assunto, problema ou com o fim de se tomar uma decisão.

[1]Universidade Federal de Mato Grosso - UFMT, Laboratório de Ambientes Virtuais Interativos – LAVI, Mato Grosso, Brasil
[2]Universidade Federal do Estado do Rio de Janeiro - UNIRIO, Rio de Janeiro, Brasil
[3]CIBERDEM – Núcleo de Ciberdemocracia
[4]Pontifícia Universidade Católica do Rio de Janeiro – PUC-Rio -Rio de Janeiro, Brasil

cmaciel@ufmt.br,claudia.cappelli@uniriotec.br,cslaviero@inf.puc-rio.br

Muitos países têm adotado diferentes formas para promover a participação dos cidadãos com o intuito de envolvê-los cada vez mais na elaboração e manutenção de suas políticas públicas (Rowe; Frewer, 2005). No Brasil, o desenvolvimento tradicional de e-Democracia tem seguido um modelo relativamente previsível e tradicional: primeiro as organizações oferecem informação, em seguida adicionam serviços e, então, iniciam a tentativa de adicionar ferramentas interativas para participação, o que não tem se mostrado um caminho muito interessante dado os baixos índices de participação apresentados num relatório global bianual sobre governo eletrônico (UNPAN, 2014). Neste, o Brasil não figura entre os países de maior pontuação em nenhum dos três níveis de participação (*e-Information, e-Consultation, e-Decisionmaking*), inclusive perdendo posição para outros países da América Latina e do mundo. O relatório evidencia que a participação eletrônica merece atenção, e que novos mecanismos para incentivar a participação, com consulta e deliberação no ambiente Web, devem ser disponibilizados.

A Sociedade Brasileira de Computação (SBC), por sua vez, definiu os cinco desafios da computação brasileira para o decênio de 2006 a 2016 (SBC, 2006). Entre estes desafios, o "Acesso participativo e universal do cidadão brasileiro ao conhecimento" (desafio 4) trata do vencimento das diversas barreiras tecnológicas, educacionais, culturais, sociais e econômicas com vistas a endereçar a questão do acesso ao cidadão brasileiro ao conhecimento. Como os proponentes do desafio descrevem, este acesso não diz respeito somente ao acesso a informações mas também a motivação de seu uso e estudo por parte dos cidadãos. Tais questões são pertinentes ao contexto de governo eletrônico e em particular, como aqui sugerido, ao contexto de participação eletrônica (e-participação).

Buscando uma definição específica para "acesso participativo" e "conhecimento" no contexto democrático, o primeiro se traduz na utilização das TIC para informação, consulta e deliberação, enquanto o segundo se traduz nos produtos da utilização destes ambientes, como a ciência acerca de questões políticas em discussão, que tornam o cidadão consciente tanto de direitos e deveres discutidos por seus representantes nas diferentes esferas governamentais.

No entanto, o desenvolvimento de sistemas deliberativos é complexo do ponto de vista computacional, bem como quaisquer outros sistemas que almejem o acesso participativo. Em termos de infraestrutura, vê-se a necessidade, por exemplo, de garantia de qualidade na entrega de informações e serviços aos cidadãos. Na perspectiva do software, o desenvolvimento de aplicações regidas por conceitos democráticos também é um desafio, dado que o acesso universal neste tipo de aplicação torna-se característica fundamental para que esta atinja seus objetivos. Além disso, podem ser discutidas questões relacionadas ao engajamento dos cidadãos em tais questões, uma vez que a introdução da tecnologia tem potencial para influenciar uma cultura,

que neste caso é a da participação democrática dos cidadãos. E, para que o cidadão possa se engajar, precisa estar informado, tendo fácil e transparente acesso à informações governamentais. Tais questões carecem de estudos, haja vista o desafio de respeito as diferenças sociais e culturais, especialmente em um país de dimensões continentais como o Brasil cujas diferenças variam enormemente. Estas e outras questões são compartilhadas com aquelas descritas no desafio do acesso participativo e universal do cidadão brasileiro, e se tornam críticas neste tema, uma vez que o conjunto de questões políticas e governamentais a qual propomos acesso participativo e conhecimento, moldam e estruturam a sociedade na qual estamos inseridos.

Cabe salientar ainda que, em 2012, a Comissão Especial de Interação Humano- Computador (CEIHC) da SBC promoveu o GranDIHCBR que, inspirado no Seminário dos Grandes Desafios da Pesquisa em Computação no Brasil (SBC, 2006), prospectou questões de pesquisa na área de IHC que serão importantes para a ciência e o país em um período de 10 anos (2012-2022), estendendo o alcance do Desafio 4 da SBC. Os resultados desse evento geraram um relatório técnico (Baranauskas et al., 2014). Nesse, as propostas apresentadas foram agrupadas em 5 grupos temáticos, a saber: Futuro, Cidades Inteligentes e Sustentabilidade; Acessibilidade e Inclusão Digital; Ubiquidade, Múltiplos Dispositivos e Tangibilidade; Valores Humanos e Formação em IHC e Mercado. As questões discutidas em cada um desses grupos são desafiadoras também no campo da e-participação e merecem atenção.

Face ao exposto, entende-se que o desenvolvimento de ambientes que promovam a e-Participação como um desafio que requer arquiteturas e metodologias mais robustas e abrangentes, além da construção de uma infraestrutura efetiva que estimule e suporte a participação do cidadão. Também, o incentivo a uma Educação que conduza os cidadãos a participarem da vida pública, com transparência nas informações e engajamento mediado pelas tecnologias, é um desafio. Frente a estas questões, propõe-se a discussão e aprofundamento das pesquisas e práticas deste tema como parte dos Grandes Desafios da Sociedade Brasileira de Computação.

E-DEMOCRACIA E E-PARTICIPAÇÃO

Embora com definições diferentes acerca do tema (Islam, 2008), muitas destas agregando e-Participação e e-Democracia (Norris, 2010) , entende-se por e-Participação como o uso das TIC para promover o engajamento dos cidadãos, tornando-os participantes ativos em decisões que influenciem a eles próprios e a sociedade que os cerca, em um nível mais avançado de relacionamento entre cidadãos e o governo que os níveis de informação e consulta. Sendo assim, e-Participação insere-se como parte na construção de uma e-Democracia, que por sua vez também abrange a disponibilização de informação e a consulta aos cidadãos, estes inclusive considerados aspectos iniciais da e-Democracia (Islam, 2008).

Muitos países têm adotado diferentes formas para promover a participação dos cidadãos na tomada de decisões, optando por referendos, audiências públicas, pesquisas de opinião pública, negociação de regras, conferência para consensos, painéis ou júris populares, comitês consultivos públicos ou grupos focais (Rowe; Frewer, 2000). No Brasil, o regime democrático é, na essência, representativo, uma vez que por eleições são definidos os governantes. Como formas de manifestação direta da soberania popular existentes na Constituição da República Federativa do Brasil, há os referendos, os plebiscitos e as iniciativas populares. Outras instâncias governamentais em segmentos distintos, como, por exemplo, os colegiados na educação, possuem pequenos grupos de representantes, selecionados pelo grupo maior que decidem sobre determinados assuntos (Maciel, 2008).

A participação envolve um relacionamento entre o governo e cidadão (Maciel, 2008). Primordialmente tal relacionamento, parte da relação governo-cidadão, é realizado através de *e-mails*, *chats* ou fóruns de discussão (Maciel, 2008; Phang; Kankanhalli, 2008; Tambouris *et al.*, 2007). A deliberação, por sua vez e em sua maioria é restrita a consultas, através de enquetes. Algumas ferramentas específicas são encontradas na Internet, em nível internacional, mas, embora façam uso de recursos integrados, não possuem foco no cidadão como indivíduo responsável pela tomada de decisão coletiva.

Aplicações com fins consultivos e deliberativos governamentais, apesar de forte uso, apresentam ainda alguns problemas uma vez que em geral (Maciel *et al.*, 2011): a) não se apresentam como um espaço de sociabilidade, com foco no cidadão como indivíduo responsável pelas decisões comunitárias, b) devem permanecer em aberto em um intervalo de tempo e ter uso efetivo, c) carecem de mecanismos estruturantes nas discussões, d) não propiciam a deliberação e, consequentemente, não viabilizam a tomada de decisões conjunta entre governo e cidadão, e f) não permitem verificar se há maturidade na participação dos indivíduos sobre as temáticas em discussão, proporcionando uma deliberação consciente.

A democracia direta, apoiada pelas TIC, tenta encontrar uma solução para a falta de participação direta dos cidadãos na tomada de decisões. A *Organisation for Economic Co-Operation and Development* (OECD, 2006) identificou cinco desafios para a e-democracia: problema de escala (tornar-se disponível para todos); capacitação e construção da cidadania; garantia de coerência das informações; avaliação da efetividade do processo; e garantia de continuidade do processo. A organização reforça ainda que os fatores críticos para o desenvolvimento de ferramentas de e-democracia, e subsequente adoção por parte dos cidadãos, incluem acessibilidade, usabilidade e segurança.

Cabe salientar que embora estas TIC possibilitem a troca de informações entre cidadãos, é notado que uma só TIC não consegue cumprir todos os objetivos de um

processo participativo (Phang; Kankanhalli, 2008), sendo necessária a construção de uma arquitetura que possa propiciar a e-Participação,integrando uma ou mais TIC. Isto pode ser percebido, por exemplo, em ambientes como o Portal da Participação Popular (SPGPC, 2012), em que são utilizadas diferentes ferramentas para disponibilizar informações, consulta e votação; ou no portal de e-Democracia, da Câmara dos Deputados (CD, 2012), que utiliza chats e fóruns para permitir a troca de opiniões entre cidadãos em diferentes níveis de engajamento.

PESQUISAS EM E-PARTICIPAÇÃO

As pesquisas em e-Participação tendem a ser multidisciplinares, envolvendo áreas tais quais antropologia, sociologia, psicologia, comunicação, tecnologia, entre outras (Araujo *et al.*, 2011)(Macintosh *et al.*, 2009). Esta multidisciplinaridade tem por razão a necessidade de alinhar o domínio de participação dos cidadãos com as tecnologias para implementá-los (Phang; Kankanhalli, 2008)(Tambouris *et al.*, 2007). Segundo Macintosh *et al.* (2009), os principais desafios de pesquisa tem sido:

- a profundidade de pesquisa: fragmentação de pesquisas e pouca persistência em abordagens;
- o design da pesquisa: baixa qualidade metodológica;
- design tecnológico: falta de formas de representação e análise de dados da e-Participação; e
- resistência das instituições: necessidade de pesquisas sobre as necessidades de avaliar os motivos de baixa adoção da e-Participação.

Segundo Shorr e Stolfo (1998), grupos de investigadores de diversas áreas (computação, estatísticas e ciências sociais), devotados aos sistemas de informação governamentais, devem realizar pesquisas aplicadas e incentivar a formação de estudantes com essa visão, a fim de estabelecer rumos de excelência às aplicações governamentais. Como desafios da democracia, Pratchett (2007) ressalta que, apesar de atualmente mais países serem democráticos, a democracia representativa nunca foi tão contestada e a democracia participativa apresenta problemas de maturidade e engajamento de cidadãos.

O design tecnológico é também destacado por outros pesquisadores como Saebo *et al.* (2008) e Velikanov (2010). Este segundo comenta que podemos esperar uma participação muito maior se fornecermos as ferramentas apropriadas para deliberar pela internet de forma eficiente e produtiva. Phang e Kankanhali (2008) evidenciam que para ampliar o sucesso da e-Participação, além da definição do objetivo a ser atingido na e-Participação, é fundamental a seleção das técnicas e ferramentas corretas, além de

também comentar sobre a necessidade de pesquisas em relação ao acesso universal, de modo a evitar a "divisão digital", formada pela falta de adequação das TIC para cidadãos com perfis distintos. Saebo *et al.* (2008) também realizam uma investigação na área de e-Participação, exaltando a falta de definições e análises sobre áreas de pesquisas e métodos, além da falta de limites sobre o que é ou não da área. Os autores propõem uma agenda de pesquisa, com seis itens: normativa, instrumental, descritiva, avaliativa, tecnológica, teórica e metodológica, compreendendo itens em que pesquisas são escassas ou carecem de testes.

Com relação à organização da área de e-Participação, diversos pontos de vista podem ser considerados. Wimmer (2007) propõe análise da e-Participação sobre quatro aspectos: estágios na formulação de políticas, *stakeholders*, nível de engajamento e áreas de participação. Os estágios na formulação de políticas dizem respeito aos passos realizados para implementar uma determinada política, e envolvem desde a criação de uma agenda política até a sua avaliação. Os *stakeholders* são aqueles interessados na participação, e que por sua vez podem ter diferentes objetivos durante a participação. O nível de engajamento também pode influenciar a condução do processo de participação. Por fim, a área de participação pode influenciar os resultados desejados na participação. A Figura 1 ilustra a relação entre estes quatro pontos.

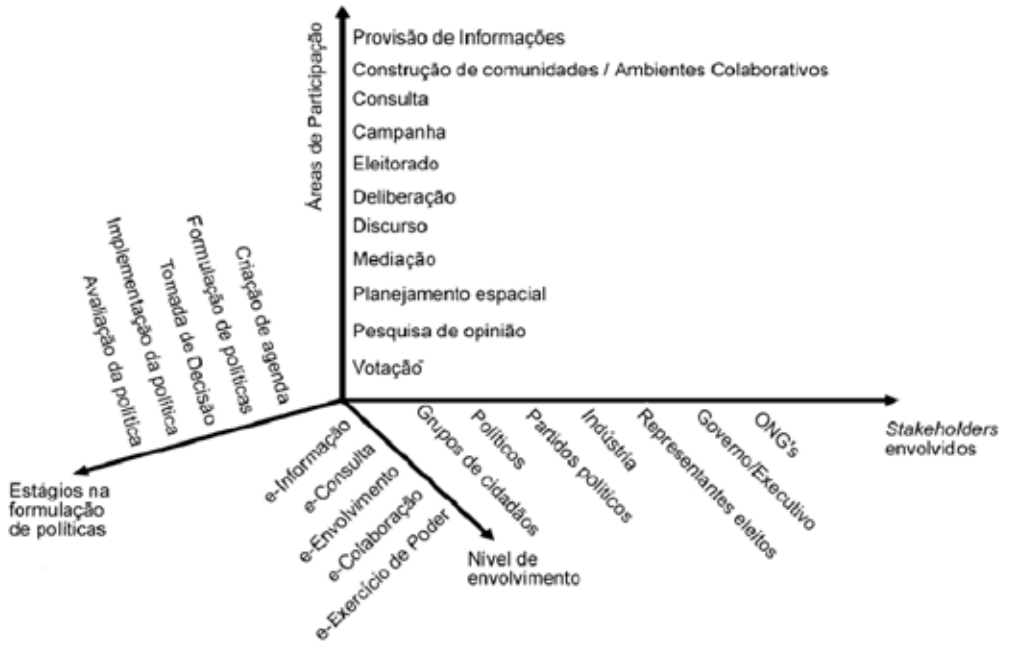


Figura 1. Rascunho de framework para caracterizar pesquisa e aplicação da e- Participação [Adaptado de Wimmer(2007) por Slaviero (2012)]

Já Staiou e Gouscos (2010) apresentam dois pontos de vista da área de e- Participação: de acordo com a forma de atividade e interação (participação na garantia de políticas, na promoção de políticas e participação social) e com o envolvimento (nível de informação consulta e participação).

No que se refere a integração de processos consultivos e deliberativos, Maciel (2008) propôs o Modelo Interativo Governo-Cidadão, organizado em fases, sendo o debate estruturado por meio da Linguagem de Interação Democrática, a DemIL, e os cidadãos sociabilizados em uma comunidade virtual. Com vistas a diagnosticar a efetividade da participação dos cidadãos em processos com fins e-democráticos, apresenta-se o método Maturidade na Tomada de Decisão – MTD. Esse método constitui-se de um conjunto de indicadores que permite monitoramento de uso e consequente medição da atuação do cidadão, desde que esse mostra seu interesse em participar do processo deliberativo, realizando seu cadastro, participando de discussões e de votações, atuando em um ambiente de sociabilização e consultando uma biblioteca de informações. A tomada de decisão dos cidadãos é classificada em imatura, pouco madura, madura ou socialmente madura.

Slaviero (2012) propôs um método, baseado em ontologia, para auxiliar a modelagem de ambientes de e-Participação por projetistas. Como resultado desse estudo, é proposta uma ontologia sobre e-Participacão denominada ePDO (*e-Participation Domain Ontology*) para relacionar as diferentes formas de participação e as tecnologias de informação e comunicação e, mais que isso, um método, denominado ePEEM (*e-Participation Environment Elaboration Method*) para elaboração de ambientes e-Participativos a partir desta ontologia. Este método foi avaliado em um projeto piloto eos resultados são animadores quanto a utilidade do método ePEEM via ontologia, mostrando que os projetistas responsáveis pela modelagem, ao obterem maior conhecimento do domínio, foram capazes de elaborar ambientes de e- Participação condizentes com o cenário proposto.

Diirr (2011) apresenta uma proposta que recria, em um ambiente virtual, as conversas comuns em diferentes contextos públicos, possibilitando o compartilhamento deexperiências, opiniões, sugestões e explicações a respeito do processo que descreve o serviço público prestado, com objetivo de aproximar Cidadãos e prestadores de serviços, além da identificação de oportunidades de melhoria para o processo. A aplicabilidade da proposta foi avaliada através de estudos de caso, onde se verificou a convergência das conversas para o serviço em discussão, o estímulo à interação entre os envolvidos com o serviço público, além da identificação de melhorias nos serviços prestados a partir das manifestações feitas pela Sociedade.

Mais recentemente, com o uso intenso das redes sociais, outras pesquisas também tem se apresentado sobre a discussão do uso destas redes no processo de participa-

ção popular. Muriana *et al.* (2013), apresentaram um estudo sobre o uso massivo do Facebook durante as manifestações em 2013, uma vez que com o uso da rede social, ficou simples e rápido para as comunidades físicas e virtuais divulgarem e organizarem os manifestos. A mobilização dessas pessoas em prol de assuntos de interesse comum, facultada pela tecnologia, permitiu-lhes arregimentar informações à inteligência coletiva. Os autores analisaram a interação entre os usuários, as páginas de eventos de manifestações e *fanpages*, porém o volume de informações ali colocado e a forma descritiva e sequencial ainda é um grande desafio para o tratamento das mesmas. Já França e Oliveira (2014) realizaram uma análise dos sentimentos da população brasileira acerca dos mesmos protestos. Por meio de uma base criada com *tweets* escritos em português brasileiro, foram pré-processados os corpus de mensagens com menos ruídos. Esse corpus foi analisado para extração do sentimento presente nas mensagens. Os autores observaram a polaridade (apoio ou repúdio aos protestos) expressa nos *tweets* e concluíram que, de acordo com os dados analisados, a maioria das mensagens apoiou os protestos.

Outro ponto importante dentro do tema de redes sociais e e-Participação é a veracidade das informações. Como todo e qualquer cidadão pode fazer uso das redes sociais e com isso publicar informações, levanta-se o problema de que as redes sociais deveriam possuir mecanismos para garantir a veracidade destas informações. Pinheiro *et al.* (2014) propõem um forma de prover aos cidadãos mecanismos de auditabilidade de informações em redes sociais. A intenção é discutir a necessidade da verificação e validação dos conteúdos expostos por qualquer cidadão e prover aos demais que acessa esta informação maneiras de verificar a veracidade dos fatos ali colocados. Para isso é especificado um conjunto características, operacionalizações e mecanismos capazes de subsidiar o desenvolvimento das funcionalidades de uma rede social de modo que ela possa conter tais mecanismos de auditabilidade.

Há ainda um ponto que se torna fundamental para que a e-Participação possa de fato acontecer que é a Transparência de Processos e Informações (Cappelli, 2009). Para que um cidadão possa de fato participar faz-se necessário que antes este tenha acesso às informações, possa fazer uso das mesmas, que estas tenham qualidade, que ele consiga de fato entendê-las e as possa auditar. Só assim poderá ter uma participação consciente. Dentre todos estes pontos destaca-se fortemente o entendimento dado que é este que propicia a construção do conhecimento. Neste ponto o trabalho de Engiel (2012) propõe uma abordagem para melhoria do entendimento de modelos de processo de prestação de serviços públicos visando o cidadão. A abordagem compreende a definição de um catálogo contendo características, operacionalizações e mecanismos que podem ser aplicados aos modelos de processo de modo a contribuir para seu entendimento e uma sistemática para sua aplicação por analistas de processos.

EXPERIÊNCIAS EM E-PARTICIPAÇÃO

Diversas iniciativas que visam engajar os cidadãos tem se tornado cada vez mais presentes, tanto internacionalmente quanto nacionalmente. Separamos iniciativas discutidas na literatura, em ambos os níveis, a seguir.

Internacionais

Ao longo dos anos, em pesquisa na literatura da área, alguns projetos e aplicações foram criadas em nível internacional, algumas das quais comentamos nesta seção.

O projetoYouGov (Coleman, 2003 apud Maciel, 2008) realiza uma comparação entre opiniões de telespectadores de reality-shows e pessoas predispostas politicamente, averiguando se o interesse em participar nas duas esferas é similar. O projeto DUCSAI (Monnoyer-Smith, 2004 apud Maciel 2008) relata debates realizados entre cidadãos sobre a localização de um novo aeroporto em Paris. Mostra que cidadãos podem participar completamente de decisões estratégicas e políticas. Kavanaugh *et al.* (2005 apud Maciel, 2008) descrevem um modelo online de participação dos cidadãos, com o uso de redes sociais, e modificação de ferramentas utilizadas para discussão e votação.O projeto Webocracy (Mach *et al.*, 2003 apud Maciel, 2008) propõe uma ferramenta para suporte eficiente na troca de informações entre cidadãos, fornecendo módulos para discussão, enquetes, gerenciamento de conteúdo Web e ajuda do cidadão. O DEMOS (Wornex, 2002 apud Maciel, 2008) é um ambiente para gerenciamento de discussões e tomadas de decisão via internet. O Smartocracy (Rodriguez *et al.*, 2007 apud Maciel, 2008) é um sistema social, que permite utilizar os dados inseridos pelos cidadãos para analisar as tomadas de decisão.

Embora exemplos interessantes, alguns destes não se encontram disponíveis na internet. Hoje em dia, outros exemplos podem ser citados. Um exemplo é o conjunto de ferramentas criado pela Delib.net (Delib, 2012) que podem ser utilizadas para criar as mais variadas formas de participação. Estas ferramentas são as seguintes (Slaviero, 2012):

• CitizenSpace: Ferramenta que permite a criação de a divulgação de consultas de cidadãos pelo governo. Possui diversas funcionalidades, como gerenciamento de consultas, busca de consultas, ferramenta para consulta, relatórios, e a possiblidade de adicionar plug-ins para adicionar funcionalidades. É utilizado, por exemplo, pelo governo do sul da Austrália para centralizar consultas online;

• DialogApp: Ferramenta para criar espaços para discussão de ideias. Além da possibilidade de gerenciamento por parte do governo, a ferramenta serve como uma forma de discutir novas ideias, podendo adicionar tags, notas e comentários. Utilizado pelo

• Conselho da cidade de Bristol, no processo de orçamento participativo, para angariar ideias para distribuição do orçamento;

• Budget Simulator: Ferramenta que permite ao governo criar processos de orçamento participativo. O cidadão pode então propor gastos em áreas propostas pelo governo pelo governo. Resultados podem ser apresentados em forma de relatório, ou então exportados. O conselho de Warrington, na Inglaterra, utilizou esta ferramenta para recolher a opinião dos cidadãos quanto à utilização do orçamento da cidade;

• QuickConsult: Ferramenta utilizada para criar questionários online, diferente da CitizenSpace, que tem por objetivo ser um lugar para divulgar consultas. Assim como as outras ferramentas, permite extrair relatórios que podem auxiliar o governo na tomada de decisões. Utilizado pela Brigada de Incêndio de Londres.

Outra ferramenta que se propõe a auxiliar a democracia é a Gov2DemoSS (GOV2U, 2012a). O Gov2DemoSS é uma plataforma de código aberto personalizável e projetado como prova de conceito da utilização de TIC para facilitar comunicação, troca de conhecimentos, e modernização dos serviços governamentais. Permite a discussão de demandas entre cidadãos, utilizando-se de fóruns de discussão. Também permite a criação de petições. Tem-se ainda oweGov (GOV2U, 2012b) que tem por objetivo, por meiode planos de trabalho, investigar a utilização das TIC para a promoção da participação dos cidadãos. Para isso, são definidos modelos, ferramentas e cenários que permitam verificar os diversos aspectos da e-Participação.Outro projeto é o MySociety (UKCOD, 2014) que tem por objetivo criar sites que permitam o engajamento dos cidadãos. Exemplos destes são o FixMyStreet, para divulgar problemas em ruas; e o WhatDoTheyKnow, que permite aos cidadãos fazer perguntas a serem respondidas pelo governo, seguindo leis de transparência pública.

Iniciativas de e-Participação na Europa também são descritas em mais detalhes em Panopoulou *et al.* (2009). Neste trabalho os autores, a partir de um modelo, reportam um conjunto de iniciativas de e-Participação, separando-as em áreas de e-Participação. A maior parte das iniciativas de e-Participação são locais e regionais. Notou-se que o grau de utilização das iniciativas de participação varia de acordo com a área de participação. Iniciativas consultivas tendem a ter mais utilizadas que outras iniciativas. Iniciativas em áreas de planejamento espacial e deliberativas se apresentam em menor quantidade. Os autores concluem que há um crescimento nas atividades de e- Participação na Europa sendo executadas e sendo planejadas. No entanto, há diversas oportunidades para melhorias nas próprias iniciativas sendo executadas, com transferências de boas práticas e a cooperação entre regiões e países em diferentes níveis de participação.Na Alemanha, onde a participação popular é estimulada desde os

anos 70, discussões com relação a e-Participação são feitas sob várias óticas (Mambrey, 2008). O autor acrescenta que diversas questões ainda hão de ser investigadas sobre a eficiência da e-Participação, efeitos da mobilização provenientes destas iniciativas e- Participativas, e investigações sobre ferramentas e tecnologias para e-Participação.

Nacionais

O estado do Rio Grande do Sul possui um ambiente para estimular a participação dos cidadãos, o Sistema de Participação Popular e Cidadã (SPGPC, 2012). Esta participação pode ocorrer via internet, utilizando inclusive dispositivos móveis para, ou então presencialmente, através de audiências públicas. Esta iniciativa pode ser considerada híbrida, uma vez que se utiliza de canais virtuais e reais para que os cidadãos possam interagir com o governo. Entretanto, no meio virtual, não há um espaço claro para discussão de demandas, e os cidadãos acabam apenas por serem consultados, através de votação. Um dos tipos de iniciativa presente neste ambiente é do orçamento participativo digital, presentes em diversos municípios brasileiros. Destacam-se as iniciativas de orçamento participativo digital realizadas em Porto Alegre (RS) e Belo Horizonte (MG), as quais estão entre as primeiras deste tipo. Mais recentemente, iniciativas em Ipatinga (MG) e Recife (PE) também são encontradas (Matheus; Ribeiro, 2009).

O site da Câmara dos Deputados disponibilizou uma ferramenta denominada e- Democracia (CD, 2012). O objetivo desta ferramenta é incentivar a participação dos cidadãos em busca de políticas públicas mais realistas e implantáveis. Existem dois espaços principais para discussão: as Comunidades Legislativas e o Espaço Livre. No primeiro, são apresentadas e discutidas leis em andamento no Congresso Nacional. No segundo, é possível discutir novos temas que possam se tornar leis. Os meios de interação utilizados nesta ferramenta são chats, fóruns, biblioteca de informações, enquetes, entre outros. Embora seja um ambiente bem estruturado este ainda está em fase de testes e não contém alguns elementos importantes para que o próprio processo de participação seja confiável, como a falta de uma identificação segura para evitar que o esmo usuário crie múltiplas contas.

O site de Participação Popular e o site da Câmara dos Deputados podem ser considerados exemplos com um nível maior de engajamento (Slaviero, 2012). No entanto, o Brasil ainda carece de iniciativas neste nível de engajamento. Em um nível mais baixo de engajamento, as ouvidorias são exemplos mais presentes no governo brasileiro. A prefeitura Municipal de Cuiabá (Prefeitura de Cuiabá, 2011) disponibiliza uma ouvidoria online, que consiste em uma página para que os cidadãos possam apresentar denúncias, elogios, solicitações de informação, reclamações e sugestões, online ou por telefone. No entanto, este tipo de interação entre cidadão e governo não possui caráter e-Participativo, na opinião de Slaviero (2012). Cidadãos apenas expõem pro-

blemas e dúvidas, não podendo discutir sobre estes e tomar decisões, como seria em um ambiente deliberativo. Da mesma forma, a Prefeitura da cidade do Rio de Janeiro também disponibiliza um espaço online de ouvidoria (PMRJ, 2012). Em nível nacional, a Ouvidoria Geral da União foi criada para receber "[...] denúncias, reclamações, sugestões, elogios e pedidos de informação referentes a procedimentos e ações de agentes, órgãos e entidades do Poder Executivo Federal". É possível criar denúncias, identificadas ou anônimas.

Quando consideramos as iniciativas provenientes do cidadão, encontramos exemplos que permitem a discussão entre cidadãos, e embora estas iniciativas não sejam oficialmente suportadas pelo governo, alguns representantes já se encontram cadastrados nela. Um exemplo é o Cidade Democrática (Seva, 2012), um espaço para criar e discutir propostas nas cidades. Cidadãos podem expor uma determinada proposta, que pode ser comentada por outros cidadãos, ONG's, e representantes devidamente cadastrados. Os cidadãos podem se posicionar contra ou a favor das propostas. Outro exemplo de iniciativa provenientes dos cidadãos para aproximá-los das decisões é o Vote na Web (WEBCITIZEN, 2012). Nele, projetos de lei apresentados pelos representantes podem ser votados e comentados entre os usuários. No entanto, convém lembrar que estes comentários não são necessariamente considerados na votação final feitas entre os representantes dos cidadãos.

Outra forma de engajar os cidadãos, prevista em lei, é através de iniciativas públicas. Uma forma de iniciativa pública é através da criação de petições. Nelas, cidadãos são chamados a analisar uma determinada questão e podem apoiar esta questão, confirmando este apoio assinando a petição. Num meio eletrônico, esta assinatura pode ser a disponibilização do nome e/ou endereço eletrônico em uma lista pública de apoiantes de determinada petição. O site Petição Pública (PPB, 2012) é um exemplo deste tipo de estratégia realizada via internet. O site permite a qualquer cidadão criar petição, angariando assinaturas eletrônicas, via internet. No entanto, esta forma especifica de iniciativa pública pode gerar desconfiança, principalmente porque o cidadão não tem garantias de que esta petição será aceita como oficial perante o governo, uma vez que esta é uma iniciativa proveniente do cidadão. Embora a constituição brasileira não restrinja o formato da assinatura, é importante lembrar que esta não considerava a existência de meios eletrônicos de assinatura, como a internet. Segundo Slaviero (2012), pode haver também problemas com relação à confiabilidade das assinaturas, uma vez que não há, pelo menos neste site, uma garantia de que apenas um cidadão poderá assinar apenas uma vez a petição, o que seria quase impossível numa petição não-virtual.

5. PROBLEMAS DE PESQUISA

Os problemas de pesquisa com os quais buscamos contribuir se referem diretamente ao desafio 4 "Acesso participativo e universal do cidadão brasileiro ao conhecimento",

levando-se em conta principalmente a Interação Humano-Computador (IHC), o uso de Redes Sociais, a educação digital e a simplicidade na representação da informação de modo a gerar maior entendimento e consequentemente propiciar o aumento do conhecimento pelo cidadão. Nesta perspectiva, a participação eletrônica dos cidadãos pode e deve ser objeto de pesquisa e aplicação.

Na área de Engenharia de Software e IHC, de forma interdisciplinar, faz-se necessária a construção de modelos integrados de consulta e deliberação e do uso de ontologias para e-participação em diferentes dispositivos, em especial nos móveis buscando auxiliar os órgãos governamentais com a disponibilização de métodos e artefatos úteis as equipes de desenvolvimento de sistemas. Nesta área ainda, diversos órgãos governamentais tem se preocupado com o uso de tecnologia para atrair a participação dos cidadãos. O uso de tecnologias como emails e enquetes não tem surtido mais efeito. Perfis e *Fanpages* em Redes Sociais têm sido muito mais utilizados por alguns órgãos públicos. Todavia, se essa informação é utilizada ou não e que ganhos poderia trazer para efetiva participação do cidadão são questões que ficam em aberto.

Também, estudos aprofundados na área de Transparência Organizacional (Cappelli *et. al*, 2013) com a intenção de dar às organizações ferramentas que permitam a esta implementar práticas que garantam acesso, uso, qualidade, entendimento e auditabilidade de processos e informações se mostram necessários.

Um outro tema bastante importante no âmbito do acesso participativo e universal do cidadão é a Educação Digital, uma vez que esta também é base para a disseminação das informações mas principalmente para a formação do cidadão de modo que este se torne um ser cada vez mais participativo. Sem este apoio, estratégias para aproximar cidadãos dos governos por meio das tecnologias podem não resultar na efetiva participação desses na vida pública, e tão somente, no consumo de informações e serviços disponibilizados pelos órgãos públicos.

Crescem atualmente diversas iniciativas populares de divulgação de informações sobre a política brasileira, em especial com a organização destas informações e apresentação a outros cidadãos por meio de aplicativos móveis e ambientes na web. Em outras palavras, está se formando uma "cultura digital", em estagio embrionário, de acesso do cidadão ao conhecimento sobre a política e o governo brasileiro. É de competência da área de Computação estudar, tanto a emergência de tais ambientes como as influências destes sobre os cidadãos brasileiros quanto ao engajamento democrático e a geração de conhecimento a partir destes ambientes. Estudos em IHC, tanto sob perspectivas de teorias cognitivas quanto semióticas, permitem analisar quantitativamente e qualitativamente este impacto. Para além de estudos somativos, a área de IHC dispõe de artefatos que possibilitam o estudo formativo de

tais ambientes. Tais pesquisas podem gerar não somente insumos para o avanço da participação eletrônica mas também podem servir como estudos de caso que auxiliem no avanço da própria área de pesquisa, em especial no que diz respeito a aspectos culturais na interação. O Brasil tem características regionais únicas que podem ser de grande valia para o avanço em tais pesquisas.

SOLUÇÕES PROPOSTAS

Trabalhos como o de Maciel (2008) e de Slaviero (2012) precisam ter continuidade, não somente sob o ponto de vista da pesquisa, mas especialmente da aplicação de tais métodos junto a órgãos governamentais. Neste sentido, além da construção de modelos integrados de consulta e deliberação e do uso de ontologias para e-participação, tem sido realizadas pesquisas e estão em produção instrumentos para analisar a qualidade de sites governamentais que disponibilizam informação, serviços e estratégias de e- participação, por meio de trabalhos monográficos ligados ao Laboratório de Ambientes Virtuais Interativos da UFMT. Também, tem-se investigado cada vez mais questões ligadas a usabilidade, comunicabilidade e acessibilidade de sites governamentais, seja em ações voltadas a sistemas em desenvolvimento que necessitam se atentar para tais aspectos da qualidade, seja na realização de testes específicos em sistemas já desenvolvidos. Ainda, pesquisas têm sido direcionadas para o uso de diferentes dispositivos, em especial dos móveis. Desta forma, se busca auxiliar os órgãos governamentais com a disponibilização de métodos e artefatos úteis as equipes de desenvolvimento de sistemas.

A construção de um Modelo de Maturidade em Transparência Organizacional[1] que organize práticas que garantam acesso, uso, qualidade, entendimento e auditabilidade de processos e informações parece bastante adequado ao momento das organizações públicas que tem que fazer valer os princípios da transparência. Dentro deste contexto, ações de pesquisa muito fortemente ligadas ao entendimento de informações propiciando o conhecimento (Engiel, 2012), a auditabilidade (Pinheiro, 2014) e a participação (Diirr, 2012) tem sido desenvolvidas. Relatos de experiências com a característica da Transparência também tem sido realizados juntos a órgãos públicos, de modo a averiguar a participação cidadão, como por Oliveira e Maciel (2013).

Na área de Educação sugere-se, entre outros, o desenvolvimento de soluções para desenvolver a educação para a transparência e o incentivo ao uso de dados abertos governamentais na forma gráfica como meio de atrair a atenção e facilitar o entendimento dos dados pelos cidadãos (Maciel *et al.,* 2012) (Antunes *et al.*, 2014) de modo a elevar e melhorar o nível de participação democrática no país.

[1]https://sites.google.com/site/ciberdem/modelo-de-maturidade-em-transparncia- organizacional

Esta proposta consolida-se com a realização de parcerias com órgãos públicos preocupados com o retorno à sociedade, como os Tribunais de Contas dos estados ou com outros que busquem a transparência e a participação do cidadão na vida pública. Algumas iniciativas já têm sido buscadas, neste sentido, mas cremos ser necessária uma maior interação entre universidade e governo neste sentido.

PLANO DE DESENVOLVIMENTO

Para as soluções apresentadas na seção 6, pretende-se, por meio da construção de parceria com empresas de governo ou organizações com interesses em e-participação, desenvolver as seguintes ações:

Desenvolvimento e Aplicação de Modelos e Métodos de E-Participação

• Analisar modelos de e-participação (Maciel, 2008) e desenvolver novos, que possam ser aplicados na construção de softwares para esta finalidade, considerando a possibilidade de processos consultivos e/ou deliberativos;

• Desenvolver, testar e/ou aplicar sistemicamente Ontologias de e- participação (Slaviero, 2012);

• Desenvolver, testar e/ou aplicar indicadores que possam ser usados para medir o nível de e-participação do cidadão (Maciel, 2008);

• Construir e/ou revisar normas e padrões para questões ligadas a usabilidade, comunicabilidade e acessibilidade de sites governamentais voltados para e-participação;

• Desenvolver e testar aplicações em dispositivos móveis para fomentar a e-participação;

• Realizar estudos sobre processos de Orçamento Participativo que façam uso de tecnologias;

Desenvolvimento e Aplicação do Modelo de Transparência Organizacional:

• Dar continuidade a estruturação do conhecimento existente no Framework de Transparência (Cappelli *et al.*, 2013);

• Dar continuidade a construção das práticas dos níveis de maturidade em Transparência;

• Desenvolver uma Ontologia de Transparência Organizacional;

• Formalizar o Método de Avaliação em Transparência Organizacional;

• Buscar aprovação do Modelo de Maturidade de Transparência organizacional e do Método de Avaliação em Transparência Organizacional junto a órgãos de governo de modo a torná-lo o modelo/método de referência nacional neste tema;

• Criar o Selo de Transparência;

• Desenvolver padrões e indicadores de monitoramento de sites para verificação de atendimento às práticas de transparência indicadas no Modelo de Maturidade em Transparência Organizacional;

• Desenvolver modelo para auditabilidade de informações em redes de conhecimento;

• Elaborar material para divulgação e formação no Modelo de Maturidade em Transparência de modo a disseminá-lo em todo o país;

• Formar pessoas que possam ajudar organizações a aplicar o Modelo de Maturidade em Transparência e realizar avaliações com uso do Método de Avaliação em Transparência Organizacional;

Educação Digital:

• Fomentar práticas educacionais que estimulem a discussão da importância da participação dos cidadãos na vida pública;

• Desenvolver ferramentas para acompanhamento contínuo pelos cidadãos de ações de e-participação;

• Inserir conteúdo/disciplina sobre Transparência no currículo do Ensino Médio;

• Produzir material didático e práticas de ensino em Transparência e e- Participação;

• Fomentar a importância de Transparência e da e-Participação junto a alunos e professores no Ensino Médio;

• Formar professores para aplicar conteúdos/disciplina de Transparência;

• Definir indicadores para acompanhamento da efetividade do ensino de Transparência;

• Disponibilizar conteúdo sobre Transparência nas mais diversas mídias e redes de informação, em especial, fazendo uso de Dados Abertos Governamentais;

Também, de uma forma geral, outras ações são importantes para fomentarmos a participação eletrônica exitosa, tais quais:

• Promover debate e fomento às leis e normativas da participação popular pelo uso das tecnologias;

• Promover práticas de marketing, concursos, campanhas e outras estratégias que auxiliem a participação e a produção de softwares neste contexto;

• Realizar estudos junto aos cidadãos, que permitam analisar como se dá ou se pode implantar processos e-participativos com tecnologias;

• Incentivar as cooperações, discussões e publicações científicas nesta área, numa parceria entre universidades e órgãos governamentais;

• Considerar experiências e boas práticas de outros países;

• Realizar as pesquisas nesta área de forma interdisciplinar, uma vez que diversas áreas do conhecimento têm relação com a temática.

CONCLUSÃO

A participação popular é um importante eixo da nossa sociedade democrática. As tecnologias nos permitem hoje projetar diferentes formas de promover a participação popular. Todavia, precisamos de estratégias mais arrojadas e efetivas para colocar o Brasil em um cenário próspero neste nível de governo eletrônico. Acredita-se que o estreitamento de relações entre os órgãos públicos e as universidades são caminhos salutares para que tais estratégias possam ser discutidas, projetadas e implantadas por meio da pesquisa científica.

Ao longo deste texto, algumas pesquisas e experiências em e-Participação foram apresentadas. Elas se diferenciam em vários níveis de engajamento, áreas de aplicação, uso de recursos tecnológicos e tipo de propostas. Esse conjunto de possibilidades traz mais desafios à implementação da e-participação, uma vez que as estratégias precisam ser modeladas considerando vários fatores. Com a profusão das redes sociais, por exemplo, pode não ser fácil atrair os cidadãos para outros canais de discussão. Todavia, estratégia integradas, considerando ecossistemas digitais, podem surtir efeitos satisfatórios.

Pela natureza da área de governo eletrônico, as discussões têm caráter multidisciplinar, exigindo que aspectos da gestão, das leis, das ciências sociais, das tecnologias e humanos, por exemplo, sejam considerados na proposta de soluções e-participativas.

Face ao exposto, fica claro o desafio que a comunidade brasileira tem de fomentar a discussão e prática da participação eletrônica de forma organizada e mediada pelas tecnologias. Para tal, a integração entre academia e órgãos governamentais é requerida.

REFERÊNCIAS

Araujo, R.; Cappelli, C.; Diirr, B.; Engiel, P.; Tavares, R. L.;Democracia Eletrônica. In: Pimentel, M.;Fuks, H. (Eds.). Sistemas Colaborativos. 1. ed. : Elsevier, 2011. p. 110- 121.

Baranauskas, C., Souza, C. S. de; Pereira, R. (Org.). I GranDIHC-BR - Grandes Desafios de Pesquisa em Interação Humano-Computador no Brasil. Relatório Técnico. Comissão Especial de Interação Humano-Computador (CEIHC) da Sociedade Brasileira de Computação (SBC). 56 p. (2014).

CD. e-Democracia. Disponível em: <http://edemocracia.camara.gov.br/o-que-e>. Acesso em 14 de setembro de 2014.

Cappelli, C. Uma Abordagem para Transparência em Processos Organizacionais Utilizando Aspectos.Tese de Doutorado: PUC-Rio, 2009.

Cappelli, C., Engiel, P., Araujo, R.M., Leite, J.C.S.P.; Managing Transparency Guided by a Maturity Model. 3rd Global Conference on Transparency Research HEC PARIS, 2013.

DELIB. Delib - A digital Democracy Company. Disponívelem: <http://www.delib.net/>. Acessoem: 10 ago. 2012.

De Mendonça, P.G.A. ; Maciel, C.; Viterbo, J. ;Visualizing infestation in urban areas. In: Proceedings of the 15th Annual International Conference on Digital Government Research - dg.o '14. New York: ACM Press, 2014. p. 186-191.

Dirr, B. Aaraujo, R.M., Cappelli, C.; Talking about Public Service Processes. In ePart 2011: 252-261.

Engiel, P. Projetando o Entendimento de Modelos de Processos de Prestação de Serviços Públicos.Dissertação de Mestrado: Universidade Federal do Estado do Rio de Janeiro, 2012.

Franca, T.; Oliveira, J.; Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. In III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2014, Brasilia. Anais do Congresso da SociedadeBrasileira de Computação. Porto Alegre: SBC, 2014.

GOV2U. Gov2DemoSS. Disponível em: <http://www.gov2demoss.org/>. Acessoem: 10 ago. 2012a.

GOV2U. We-gov Project. Disponível em: <http://www.wegov-project.eu/>. Acessoem: 10 ago. 2012b.

Islam, M.S.;Towards a sustainable e-Participation implementation model. European Journal of ePractice, v. 5, p. 1-12, 2008.

Maciel, C. Um método para mensurar o grau de maturidade na tomada de decisão e-Democrática.Tese de Doutorado: Universidade Federal Fluminense. Niterói, 2008.

Maciel, C.; Roque, L.; Garcia, A. C. B. E-Democracy: Concepts, Experiences and Challenges. In Herrmann, P. (Ed.).Democracy in Theory and Action. 1. ed. New York: Nova Science Publishers, Inc., 2011. v. 8p. 51-92.

Maciel, C. ;Viterbo, J. ; Breitman, K.; Transparência Pública e Dados Abertos Governamentais. In: André Resende. (Org.). Governo Brasileiro no Futuro: sugestões e desafios para o Estado (2012-2022). 1ed.São Paulo: Cubzac, 2012, p. 47-56.

Macintosh, A.; Coleman, S.; Schneeberger, A. eParticipation: The Research Gaps. In First International Conference on Electronic Participation (ePart'09). Springer, 2009.

Mambrey, P.From participation to e-participation: the German case. In Anais do 2nd International Conference on Theory and Practice of Electronic Governance.: ACM. 2008.

Modelo de Maturidade em Transparência Organizacional - https://sites.google.com/site/ciberdem/modelo-de-maturidade-em-transparncia- organizacional

Muriana, L.M., Maciel, C., Garcia, A.C.B. Do Facebook às Ruas – Comunidades em Interação.In Anais do WAIHCWS 2013: 39-50.

Norris, D. F.; e-government... not e-governance... not e-democracy not now! Not Ever? In Anais do 4th International Conference on Theory and Practice of Electronic Governance - ICEGOV '10. New York, New York, USA: ACM Press, 2010.

OECD. Organization for Economic Co-operation and Development. Disponível em: <http://www.oecd.org>. Acesso em: 10 mai. 2014.

Oliveira, L.K.B., Maciel, C. Transparency and Social Control via the Citizen s Portal: A Case Study with the Use of Triangulation. In Technology-Enabled Innovation for Democracy, Government and Governance. Lecture Notes in Computer Science, v. 8061, p. 112-124, 2013.

Panopoulou, E.; Tambouris, E.; Tarabanis, K.eParticipation initiatives: How is Europe progressing. European Journal of ePractice, v. 7, n. March, p. 15–26, 2009.

Phang, C.W.; Kankanhalli, A.; A framework of ICT exploitation for e-participation initiatives.Communications of the ACM, v. 51, n. 12, p. 128, 1 dez 2008.

Pinheiro, A., Cappelli, C., Maciel, C.; Increasing Information Auditability for Social Network Users. In HCII (12) 2014, p. 536-547.

PMRJ. Sistema de Ouvidoria da Prefeitura da Cidadedo Rio de Janeiro. Disponível em: <http://www21.rio.rj.gov.br/siso/internet/ouvidoria.htm>. Acesso em: 10 ago. 2010.

PPB. Petição Pública Brasil. Disponível em: <http://www.peticaopublica.com.br/>. Acesso em: 10 ago. 2012.

Pratchett, L. Where we are: e-Consultation/e-Deliberation. In: Electronic Democracy: Achievements and Challenges, European Science Foundation - LiU Conference. Vadstena, Sweden, nov 2007. Disponível em: <http://www.docs.ifib.de/esfconference07/conf_programme.html>. Acesso em 10 jan. 2008.

Prefeitura de Cuiabá. Prefeitura de Cuiabá. Disponívelem: <http://www.cuiaba.mt.gov.br/>. Acesso em: 23 de novembro de 2011.

Rowe, G.; Frewer, L.;Public participation methods: a framework for evaluation.Science, Technology & Human Values, v. 25, Winter, p.3-29, 2000.

Rowe, G.; Frewer, L. J.; A Typology of Public Engagement Mechanisms.Science, Technology & Human Values, v. 30, n. 2, p. 251-290, 2005.

Saebo, Ø.; Rose, J.; Flack, L.S.; The shape of eParticipation: Characterizing an emerging research area. Government Information Quarterly, v. 25, n. 3, p. 400 - 428,2008.

SBC. Sociedade Brasileira de Computação. Grandes Desafios da Pesquisa em Computação no Brasil 2006 – 2016.2006. Disponível em: <http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=vie w.download&catid=50&cid=11 >. Acessoem: 10 ago. 2014.

SEVA. Cidade Democrática. Disponível em: <http://www.cidadedemocratica.org.br/>. Acesso em 14 de setembro de 2014.

Shorr, H.; Stolfo, S. J. A Digital Government for the 21st Century. Comunications of the ACM, v. 41, n. 11, p. 15-19, nov. 1998.

Silva, L.A., Magnus, S., Silveira, M.S., Maciel, C.; Designing help system for e-GOV websites: A Brazilian case study. Information Polity 18(3): 261-274 (2013).

Slaviero, C., Garcia, A.C.B., Maciel, C.; Exploiting e-Participation Using an Ontological Approach. ePart 2012a: 144-155.

Slaviero, C. Um método para modelagem de ambientes e-participativos baseado em ontologia. Dissertação (Mestrado em Computação) - Universidade Federal Fluminense, Niterói, RJ : [s.n.], 2012. 133 f.

SPGPC. Portal da Participação. Disponível em: <http://www.participa.rs.gov.br/>. Acesso em: 4 ago. 2012.

Staiou, E.R.; Gouscos, D.; Socializing E-governance: A Parallel Study of Participatory E-governance and Emerging Social Media. Integrated Series in Information Systems, v. 25, n. 3, p. 543-559, 2010.

Tambouris, E.; Liotas, N.; Taranabis, Konstantinos.; A Framework for Assessing eParticipation Projects and Tools. (Hawaii, Ed.) In Anais do 40th Hawaii International Conference on System Sciences.: IEEE, 2007.

UKCOD. mySociety. Disponível em https://www.mysociety.org/ Acesso em: 18 de agosto de 2014.

UNPAN. United Nations. UN Global EGovernment Readiness Report 2005: From E-government to E-Inclusion. United Nations Publications, 2014.Disponível em:

<http://www.unpan.org/>. Acesso em: 25 set. 2006.

Velikanov, C.; Requirements and tools for an efficient eParticipation. In Anais do 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities(dg.o), 2010.

WEBCITIZEN. VOTENAWEB Projetos de Lei. Disponível em: <http://www.votenaweb.com.br/>. Acessoem: 10 deagosto de 2012.

Wimmer, M.A.; Ontology for an e-participation virtual resource centre. In Anais do 1st International Conference on Theory and Practice of Electronic Governance - ICEGOV '07. Macau, China: ACM Press., 2007.

BIG DATA, LITTLE DATA E *BETTER DATA* EM SISTEMAS DE RECOMENDAÇÃO

Priscila M. V. Lima[1], Claudia L. R. da Motta[1], Adriano J. O. Cruz[1,2,]
Antonio J. Alencar1, Eber A. Schmitz[1,2], Jonas Knopman[1], Cabral Lima[2]

Abstract. Recommender systems have gained a widespread use in commercial activities, which is partly due to recent improvements on data collection and storage capacity, known as Big Data. Little Data, on the other hand, refers to what we know about our most usual activities. The same Technologies, such as cloud and mobile computing, are applicable both to Big and to Little Data. Big Data processing can be very complex. So, it is profitable to employ Little Data to guide that search for meaningful information, that is, to obtain Better Data. In that sense, we propose to combine recommendation systems with Big and Little data to enhance decision support in the banking/financial area.

Resumo. Sistemas de recomendação vêm sendo cada vez mais usados em atividades comerciais. Parte do seu sucesso veio do aumento da capacidade de captação e armazenamento de dados, conhecido como Big Data. Por outro lado, Little Data refere-se ao que sabemos sobre nossas atividades mais usuais. As mesmas tecnologias, como computação em nuvem e tecnologias móveis, podem ser utilizadas para Big e Little Data. Como o processamento de Big Data pode ser complexo e demorado, pode-se tirar proveito de Little Data para direcionar a busca, evoluindo-se para Better Data. Propomos combinar recomendação e Big Data com aspectos de apoio à decisão para aplicações no setor bancário/financeiro.

1. INTRODUÇÃO

Surgidos em meados dos anos noventa através dos sistemas de filtragem colaborativa, sistemas de recomendação vêm sendo cada vez mais utilizados tanto em atividades comerciais quanto sociais e até mais especificamente no meio acadêmico. Construídos a partir de estruturas de ranqueamento, tendo sua filtragem baseada em colaboração ou em conteúdo, podem empregar uma enorme variedade de técnicas tanto para representar quanto para decidir quais aspectos do domínio da aplicação devem ser focalizados e com qual importância. Dentre essas propriedades podemos citar acurácia, robustez,

[1]Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE)
[2]DCC/Instituto de Matemática
Universidade Federal do Rio de Janeiro – Rio de Janeiro, RJ – Brazil

{priscila.lima,claudiam,adriano, juarez.alencar, eber, jonas}@nce.ufrj.br, cabral.lima@ufrj.br

escalabilidade etc. Também precisam ser realizadas escolhas quanto à técnica (ou técnicas) a ser(em) utlilizada(s) na mineração dos dados, tais como: regras de associação, clusterização, árvores de decisão, *k-nearest neighbours*, análise de links, redes neurais artificiais, regressão etc. Parte do sucesso da adoção de sistemas de recomendação veio do crescimento da capacidade de captação e armazenamento de dados, mais conhecido como o fenômeno de *Big Data*.

2. *BIG DATA, LITTLE DATA*

Dados obtidos de uma série de fontes são agregados em um ambiente de armazenamento massivo, potencialmente geograficamente distribuído, para que sejam analisados a fim de que padrões sejam descobertos. Tais padrões são utilizados na tomada de decisões ao permitir previsões mais precisas, comunicações melhor direcionadas e serviços mais personalizados.

Big Data pode ser definido como o que as organizações sabem sobre pessoas, sejam elas clientes, funcionários, cidadãos ou eleitores, a partir da análise de um conjunto muito grande de informações. Bilhões de transações financeiras podem ser utilizadas para prever fraudes de cartão de crédito ou concessão de novas linhas de crédito. Já milhões de interações nas mídias sociais permitem que analistas de marketing observem novas tendências do mercado consumidor. Da mesma forma, varejistas podem combinar essas tendências com o conhecimento sobre os milhões de compras realizadas num determinado período, e assim planejar suas promoções.

Por outro lado, *Little Data* refere-se ao que sabemos sobre nós mesmos, o que adquirimos, os lugares que frequentamos, nossas atividades mais usuais. As mesmas tecnologias que possibilitaram surgimento do Big Data, tais como computação em nuvem e tecnologias móveis, também podem ser utilizadas para aumentar o autoconhecimento dos indivíduos. Entretanto, *Big* e *Little Data* diferem em três pontos básicos, resumidos na Tabela 1.

TABELA 1. Comparação entre *Big Data* e *Little Data*.

	Big Data	**Little Data**
Foco	consecussão dos objetivos governamentais	ajudar indivíduos a atingir seus objetivos
Visibilidade	Indivíduos não enxergam Big Data	ajuda indivíduos a enxergar
Controle	controlado por organizações	indivíduos concedem permissão a organizações para acessar

3. *BETTER DATA*

Tendo em conta que o processamento de *Big Data* pode ser altamente complexo e, consequentemente, demorado, um caminho que vem despontando na literatura concernente é o de tirar proveito de *Little Data* para direcionar os objetivos da busca em *Big Data*. Assim, evolui-se o conceito de *Big Data* para o de *Better* (mais relevante e computável) *Data*. Nossa proposta consiste em combinar o estado-da-arte em Sistemas de Recomendação e *Big Data* com avanços em outros aspectos de Inteligência Artificial a fim de tirar proveito do potencial de hibridização de técnicas e do potencial de parceria numa área que faz parte dos grandes desafios, [Chen et al, 2012] [Cruz et al, 2009] [Hu et al, 2012] [Magalhães and Lima, 2014] [NBUSINESS, 2014] [Ricci et al, 2010] [Wosniak et al, 2013]. O grupo de pesquisadores que constituem a equipe reúne os conhecimentos necessários para tal fim, bem como detém a experiência requerida em estudos e parcerias multidisciplinares, [Bottino et al, 2012] [Cardoso et al, 2014] [Cruz et al, 2002a, 2002b] [França et al, 2014]. Em particular, consideramos importante a incorporação da expertise em sistemas de apoio à decisão a sistemas de recomendação aplicados ao setor bancário/financeiro, [Alencar et al, 2008, 2013] [Barbosa et al, 2008] [Fernandes et al, 2014]. Ferramentas já desenvolvidas pelo grupo, como a de [Gomes et al, 2010, 2014], podem ser adaptadas para os objetivos aqui delineados.

REFERENCIAS

Alencar, A. J. S. M., Cruz, L. T., Schmitz, E. A. and Ferreira, A. L. (2008) "Using a Novel Approach to Cluster Analysis to Gain New Valuable Insights into Software-Project Risk Management". In: IEEE/IFIP International Workshop on Business-driven IT Management (BDIM), Salvador, v. 1, p. 40-48.

Alencar, A. J. S. M., Rigel P. F., Correa, A. L. and Schmitz, E. A. (2013) "Maximizing the Appropriation of the Intangible Benefits Yielded by IT Investments in the Public Sector", Journal of Software, v. 8, p. 1537-1549.

Barbosa, B. P., Schmitz, E. A. and Alencar, A. J. S. M. (2008) "Generating Software -Project Investment Policies in an Uncertain Environment". In: The IEEE Systems and Information Engineering Design Symposium (SIEDS'), Charlottesville, Virgina. Proceedings of the 2008 IEEE Systems and Information Engineering Design Symposium, v. 1, p. 30-35.

Bottino, B. and Cruz, A. J. O. (2012) A parallel method for tuning Fuzzy TSK Systems wih CUDA. In: Proceedings of SBGAMES 2012.

Cardoso, D. O., Carvalho, D. S., Alves, D. S. F., Souza, D. F. P., Carneiro, H. C. C., Pedreira, C. E., Lima, P. M. V. and FRANÇA, F. M. G. (2014) "Credit Analysis with a clustering

RAM-based neural classifier". In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges. Proc. of ESANN 2014, p. 517-522.

Chen, H, Chiang, R. H. L. and Storey, V. C. (2012) "Introduction to Business Intelligence", MIS Quarterly (MISQ) Special Issue in Business Intelligence Research, v. 36, no. 4, pp. 1165-1188.

Cruz, A. J. O., Franco, C. R. and Vidal, L. S. (2002a) "A Validity Measure for Hard and Fuzzy Clustering derived from Fisher's Linear Discriminant". In: 2002 IEEE International Conference on Fuzzy Systems, 2002, Honolulu. Procs. of the 2002 International Conference on Fuzzy Systems. Piscataway, NJ, USA, v. 1, p. 1493-1498.

Cruz, A. J. O., Raposo, R. C. T. and Mendes, S. B. T. (2002b) "Using Previous Knowledge for Stock Market Prediction Based on Fundamentalist Analysis with Fuzzy-Neural Networks". In: 2nd WSEAS International Conference on Simulation, Modelling and Optimization (ICOSMO 2002), 2002, Skiathos. Procs of the 2nd WSEAS International Conference on Simulation, Modelling and Optimization, v. 1, p. 2211-2216.

Cruz, C. C. P., Motta, C.L.R., Santoro, Fl. M. and Elia, M. F. (2009) "Applying Reputation Mechanisms in Communities of Practice: A Case Study", Journal of Universal Computer Science, v. 15, p. 1886-1906.

Fernandes, R. P., Alencar, A. J. S. M., Schmitz, E. A. and Correa, A. L. (2014) "Analysing IT Investments in the Public Sector: A Project Portfolio Approach", Journal of Software, v. 9, p. 1200-1213.

França, F. M. G., Degregorio, M., Lima, P. M. V. and Oliveira jr, W. R. (2014) "Advances in Weightless Neural Systems". In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2014, Bruges. Proceedings of ESANN 2014, p. 497-504.

Gomes, D.S.M. ; MOTTA, C.L.R. ; CRUZ, Adriano Joaquim de Oliveira . Sistema Integrado para Construção de Inferências Aplicáveis a Jogos Psicopedagógicos. Revista Brasileira de Computação Aplicada, v. 2, p. 17-32, 2010.

Gomes, D. S. M., Lima, P. M. V., Cruz, A. J. O. e Motta, C. L. R. Avaliação Empírica da Usabilidade do Sistema NÉBULA em Jogos Psicopedagógicos. RENOTE. Revista Novas Tecnologias na Educação, 2014 (aceito para publicação).

Hu, D., Zhao, J. L., Hua, Z. and Wong, M. (2012) "Network-Based Modeling and Analysis of Systemic Risk in Banking Systems", MIS Quarterly (MISQ) Special Issue in Business Intelligence Research, v. 36, no. 4.

Magalhaes, R. P. and Lima, C. (2014) "A New Method for Haar-Like Features Weight Adjustment Using Principal Component Analysis for Face Detection". In: ICONS 2014, The Ninth International Conference on Systems, 2014, NICE, France. Procs of the ICONS 2014, v. 1, p. 55-62.

NBUSINESS, (2014) "UFRJ abre inscrições para curso grátis sobre Big Data: segunda edição da Summer School on Big Data é patrocinada pela EMC e terá 150 vagas distribuídas em três minicursos", A central de whitepapers de tecnologia da COMPUTERWORLD, http://www.eventosnowdigital.com.br/computerworld/carreira/2014/04/15/ufrj-abre-inscricoes-para-curso-gratis-sobre-big-data.

Ricci, F., Rokach, L., Shapira, B. and Rantor, P. B. (Editors) (2010) "Recommender Systems Handbook", Springer, New York, USA.

Wosniak, M., Grana, M. and Corchado, M. (2013) "A survey of multiple classifier systems as hybrid systems", Information Fusion, v. 16, Elsevier, p. 3–17, March.

MONITORAMENTO E ADAPTAÇÃO DE TRANSFORMAÇÕES EM DADOS CIENTÍFICOS AO LONGO DE EXECUÇÕES PARALELAS EM AMBIENTES DE PROCESSAMENTO DE ALTO DESEMPENHO

Marta Mattoso[1] e Daniel de Oliveira[2]

Resumo. Este artigo aborda aspectos de Grandes Desafios da Computação, no Brasil, no domínio de interesse de Exploração de Petróleo. Mais especificamente, são abordadas as complexidades no tema de Gestão da Informação em grandes volumes de dados multimídia distribuídos (i.e. Big Data) em Petróleo e Gás. Essa área se caracteriza por trabalhar com um grande volume de dados binários e em formatos heterogêneos específicos do domínio. Ao longo de simulações numéricas envolvendo cálculos complexos, novos arquivos de dados científicos são gerados. O uso de computação paralela é quase que obrigatório nesse cenário. Apesar de nos últimos anos termos tido avanços significativos em termos de algoritmos e técnicas de execução paralela em alto desempenho, esse avanço não foi correspondido na análise e gerência dos dados ao longo de sua geração. Mesmo com um ambiente computacional de alto desempenho, experimentos com análises científicas em petróleo e gás costumam ficar semanas em execução. Dentro desse cenário, esse artigo aborda o problema computacional de oferecer uma solução genérica, independente de domínio, que ofereça recursos analíticos de monitoramento dessas execuções. Indo um passo adiante, são abordadas mais especificamente a dificuldade em realizar modificações em experimentos de larga escala, durante a sua execução em ambientes de computação de alto desempenho, como grades computacionaos e nuvens de computadores.

1. GRANDES DESAFIOS EM ANÁLISE DE DADOS CIENTÍFICOS

O 1º. Seminário de Grandes Desafios em Computação no Brasil, realizado em São Paulo em 2006 delineou o desafio na "Gestão da informação em grandes volumes de dados multimídia distribuídos". Naquele evento, apresentamos o potencial do apoio computacional na evolução da ciência e seus desafios ao lidar com um grande volume de dados e equipamentos computacionais com capacidade de processamento de alto desempenho com larga escala de tarefas (SBC, 2006). Ao longo desses oito anos, muitos avanços ocorreram no apoio computacional para o desenvolvimento de ciência em larga escala, chamada na europa de *e-Science* e nos EUA de *cyberinfraestructure*. Uma característica

[1]Programa de Engenharia de Sistemas e Computação (PESC)
COPPE – Universidade Federal do Rio de Janeiro (UFRJ)
[2]Instituto de Computação – Universidade Federal Fluminense (IC/UFF)

marta@cos.ufrj.br, danielcmo@ic.uff.br

importante no desafio do apoio computacional ao desenvolvimento de ciência em larga escala é a interdisciplinaridade envolvendo as diversas áreas da ciência da computação, com forte ênfase em: bancos de dados, sistemas de informação, computação paralela, redes, engenharia de *software*, interação humano-computador, computação gráfica, *Web* semântica, etc. Teorias e algoritmos clássicos da computação são tipicamente voltados para a área de negócios e precisam ser repensados para a *e-Science*. Algumas áreas incipientes há dez anos, estão hoje mais consolidadas, como por exemplo: *workflows* científicos, proveniência de dados e consulta a dados científicos, todas quase sempre associadas ao processamento paralelo. Vários eventos foram criados nesses últimos dez anos e vêm se consolidando na apresentação de contribuições em computação para esses desafios, como por exemplo:

- IEEE *e-Science Conference* (iniciado em 2004)
- IPAW (*International Provenance and Annotation Workshop*, iniciado em 2006)
- ACM/IEEE *Workshop* on *Workflows in Support of Large-Scale Science*, iniciado em 2003)
- BrESci, da SBC, o qual fizemos parte de sua idealização e organização desde 2007, (http://www.cos.ufrj.br/~marta/E-Science/)

Além desses, vários outros mais recentes foram especializados em computação em nuvem, como o *International Workshop on Clouds and (e-Science) Applications Management* (CloudAM – iniciado em 2012) e o *International Workshop on Cloud Computing and Scientific Applications* (CCSA – iniciado em 2011). O mesmo ocorreu com alguns periódicos de prestígio, da área de computação paralela, que se voltaram à publicação de contribuições dentro desse tema, mudando seus nomes para destacar a e-Science, como foi o caso do periódico "*Future Generation Computer Systems: The International Journal of Grid Computing and eScience*", da Elsevier classificado como A2 no qualis da CAPES.

Especificamente em nossos grupos de pesquisas, vimos perseguindo esses desafios nos últimos dez anos, obtendo sucesso em inovações teóricas e práticas no apoio computacional ao desenvolvimento de ciência em larga escala, com uso em aplicações reais. Dentre as contribuições teóricas, destaca-se a álgebra de workflows centrada em dados, que se apoia na álgebra relacional (Ogasawara et al. 2011). Essa álgebra proporcionou um apoio original ao fluxo de transformação de dados ao longo de simulações computacionais, que levou a soluções genéricas e otimizáveis do ponto de vista de execução paralela. Outro benefício foi o de proporcionar a submissão de consultas a esse fluxo de dados em tempo de execução (Costa et al. 2013, Santos et al. 2013). Modelos teóricos de escalabilidade, em execução paralela de workflows, com elasticidade de recursos em nuvens computacionais também foram desenvolvidos (Coutinho et al. 2014, Oliveira et al. 2012, 2013).

Do ponto de vista prático, sistemas de código aberto que incorporam as inovações teóricas desenvolvidas estão em operação foram disponibilizados em portais de *software* livre, como o Chiron (Ogasawara *et al.* 2013) e SciCumulus (Oliveira *et al.* 2010). Todas as contribuições desenvolvidas foram validadas e exploradas em aplicações científicas reais, seja da área de exploração de petróleo, seja da área de bioinformática (*e.g.* análises filogenéticas, farmacofilogenônica, *etc.*).

No nosso caso especificamente, ao trabalharmos junto a cientistas envolvidos em áreas de domínio de interesse junto à exploração de petróleo e com outros cientistas na área de bioinformática, pudemos aplicar soluções computacionalmente inovadoras em cenários de resolução de problemas brasileiros dentro dessas áreas, como por exemplo, aplicações envolvendo a exploração de petróleo em águas ultra-profundas (caso bem particular do Brasil, uma vez que a exploração da camada do pré-sal é uma prioridade) ou aplicações de bioinformática no desenvolvimento de novos fármacos para doenças negligenciadas, como é o caso de tradicionais doenças tropicais como a Malária (caso também do Brasil).

Em todos esses cenários reais existe um grande volume de dados que deve ser não somente processado, mas avaliado e auditado por terceiros. Realizar essa avaliação, onde todo o histórico de transformações deve ser registrado, em ambientes de alto desempenho não é uma tarefa trivial. Nesse sentido, embora a contribuição das técnicas computacionais desenvolvidas não seja específica para problemas brasileiros, elas puderam ser aplicadas em problemas importantes do país, muito por conta do trabalho multidisciplinar e interesse comum em problemas brasileiros tanto entre os pesquisadores da computação quanto os das áreas de aplicação.

A aproximação com cientistas que precisam trabalhar com grandes volumes de dados evidenciou os problemas em aberto na literatura, mais especificamente o de monitoramento e adaptação de *workflows*, conforme levantamento realizado por Mattoso *et al.* (2013). À medida que cientistas usam soluções de *workflows* e proveniência, novos desafios para as teorias e *softwares* sendo desenvolvidos se apresentam.

Especificamente, visamos a discussão de aspectos da interação do cientista com a execução de seu experimento computacional executado em paralelo em embientes de alto desempenho, o que é chamado de envolver o ser humano no processo (do inglês HIL: *human-in-the-loop*), um dos desafios apresentados em Jagadish *et al.* (2014) ao discutirem os desafios na gerência de grandes volumes de dados. A importância desta interação, amplamente discutida em Jagadish *et al.* (2014), torna-se imperativa nas áreas de geologia e geofísica envolvidas na exploração e produção de petróleo.

2. *WORKFLOWS* CIENTÍFICOS E DADOS DE PROVENIÊNCIA

Pretendemos trabalhar nos desafios ao proporcionar ao cientista recursos de monitoramento e adaptações na configuração de experimentos científicos que envolvolvem o processamento de grandes volumes de dados por meio de sistemas de gerência de *workflows* científicos e do apoio aos dados de proveniência.

Os experimentos científicos podem ser modelados como uma sequência de passos onde cada passo está relacionado a uma determinada simulação computacional. A essa modelagem atribuímos o nome de *workflow*. Toda essa adaptação e monitoramento devem ser realizados à medida que os dados envolvidos são transformados por programas de simulação computacional ao longo do *workflow*.

Sistemas de Gerência de *Workflows* Científicos (SGWfC) podem ser utilizados para modelar e orquestrar a execução paralela de seus experimentos em larga escala (Deelman *et al*. 2009). Muitos avanços teóricos e práticos foram realizados no desenvolvimento de sistemas de *workflows* e a gerência de proveniência dos dados científicos. Pode-se associar, ao *workflow* científico, a proveniência de dados (Freire *et al*. 2008), a qual diz respeito à composição do *workflow*: as atividades, suas características e os fluxos de dados entre elas; e ao histórico da execução do *workflow*, como, por exemplo, o tempo de execução de cada atividade e o recurso computacional responsável por sua execução (Dias 2013).

O registro estruturado de dados de proveniência é essencial para a confiabilidade e reprodutibilidade de um experimento científico. A proveniência descreve da história de geração de dados do *workflow*, o seu *pedigree* (Freire *et al*. 2008). O repositório de dados de proveniência figura como o componente central de um sistema de *workflows*. Neste repositório são armazenadas todas as informações referentes à definição e à execução do *workflow*, ou seja, proveniência prospectiva e retrospectiva. A importância do registro de dados de proveniência acabou gerando um esforço de padronização junto ao W3C, denominado PROV (Moreau *et al*. 2011), que permite a representação de entidades, agentes e atividades envolvidos na geração de um dado e seus relacionamentos.

No entanto, esse apoio ainda é voltado para gerir o encadeamento e execução paralela dos programas e nem tanto para apoiar a análise do fluxo de dados científicos sendo gerados. Dados científicos são tipicamente gerados e armazenados em arquivos binários ou de formato proprietário cujos conteúdos não estão representados nos dados de proveniência. Os dados científicos acabam por ficar isolados sem uma representação que os relacione para análise do fluxo. Uma das maiores dificuldades está em acompanhar a transformação de dados em larga escala ao longo da execução paralela do *workflow*.

Muitos avanços foram obtidos na gerência do paralelismo de dados envolvidos no fluxo de transformação de dados em sistemas de workflows científicos (Deelman et al. 2007, Fahringer et al. 2005, Ogasawara et al. 2013, Oliveira et al. 2010, Taylor et al. 2007, Wozniak et al. 2013). Tais avanços permitiram que análises de problemas científicos em escalas maiores pudessem ser realizadas com mais rapidez, confiança, qualidade e de forma reproduzível. A geração de dados de proveniência oferecida pelos SGWfC provê um apoio fundamental para analisar o histórico de transformações dos dados após a execução do workflow. Entretanto, uma crítica frequente a essa abordagem de automação do workflow científico como um todo é a execução como uma "caixa preta", sem acesso ao fluxo de transformações dos dados durante a execução. Mesmo quando executadas em paralelo em ambientes de processamento de alto desempenho, essas aplicações científicas costumam executar durante semanas e até meses, como é o caso de imageamento geofísico (Hanzich et al. 2014). Esse e outros problemas são discutidos na próxima seção.

3. GERÊNCIA DE DADOS EM *WORKFLOWS* CIENTÍFICOS EM PETRÓLEO E GÁS

O grande problema ao configurar um experimento científico é que isso exige a definição de metodologias, algoritmos, *softwares*, parâmetros, que juntos geram um número de combinações muito elevado. Fica muito difícil prever de antemão qual seria a melhor configuração a ser utilizada. Assim, o cientista seleciona uma possível combinação de programas, valores de parâmetros e algoritmos e tenta rastrear resultados parciais do experimento para verificar se a configuração escolhida precisa ser alterada, pois esperar dias por seu término para descobrir que a configuração não foi adequada não é uma alternativa plausível.

Por conta disso, alguns *softwares* de apoio à execução paralela de simulações computacionais oferecem recursos de monitoramento e adaptação da configuração dos experimentos, como é o caso de (Hanzich *et al.* 2014, Adams *et al.* 2010 e Gannon *et al.* 2007). Porém, esses *softwares* são específicos para um domínio restrito de interesse. Isso faz com que um experimento científico envolvendo diversas áreas de aplicação tenha que interagir com *softwares* distintos de gerência do experimento.

Por exemplo, na área de exploração e produção de petróleo, um sistema de *workflows* seria usado para o imageamento geofísico, outro sistema para a análise de fadiga em plataformas *offshore* e um terceiro para análises envolvendo a quantificação de incertezas em parâmetros desses mesmos *workflows*. Essas abordagens específicas a um domínio restrito causam a repetição de diversos códigos que seriam comuns a outras aplicações. Foi justamente com a motivação de evitar essa proliferação de códigos e ausência de padrões que surgiram os sistemas de *workflows* científicos e o padrão de proveniência PROV citado anteriormente. Sendo assim, a adoção de inúmeros sistemas

que possuem muitas características em comum implicam a perda de produtividade da equipe de cientistas por conta da curva de aprendizado de cada um desses novos sistemas. Por outro lado, o poder de adaptação encontrado de modo pré-programado para um domínio específico em sistemas como o BSIT (Hanzich *et al.* 2014), precisa ser generalizado e incorporado à especificação e ao modelo de execução dos *workflows* científicos nos sistemas de *workflow*.

Os *workflows* podem passar por uma série de alterações na sua definição e nos seus parâmetros para que se atinja a configuração que produza o resultado desejado para o experimento. Parte destas alterações é realizada mediante a análise dos resultados produzidos ao longo da execução do *workflow*. Um cenário muito comum é aquele em que o cientista analisa o resultado de uma primeira execução do *workflow* e então toma decisões sobre o que será processado em seguida. Ou seja, de acordo com algum resultado (seja ele intermediário ou final), o cientista decide o que deverá ser ajustado na execução corrente do *workflow* científico ou em futuras execuções do mesmo. Parâmetros e dados de entrada são alterados até que o objetivo da execução do *workflow* seja alcançado. Os ajustes mais comuns são relativos aos dados de entrada e os parâmetros.

Os cientistas comumente exploram diferentes conjuntos de dados e parâmetros e executam o *workflow* repetidas vezes para descobrir qual a melhor configuração. Entretanto, não são apenas os parâmetros e os dados que podem ser ajustados e variados. Muitas vezes os cientistas, baseados nos resultados de execuções anteriores, realizam ajustes na própria definição do *workflow*, alterando os programas associados às suas atividades a fim de obter um desempenho melhor, um resultado com mais qualidade ou mesmo um programa mais adequado ao perfil dos dados que estão sendo gerados.

Pode ocorrer o caso em que uma mesma atividade do *workflow* tenha várias implementações (programas) correspondentes. Em simulações numéricas, por exemplo, uma atividade de solução de sistemas de equações lineares pode utilizar diferentes metodologias, métodos, algoritmos e programas correspondentes. Assim, estes programas podem ser vistos como alternativos entre si e, de acordo com o ambiente de execução e com os requisitos do experimento, um programa pode ser mais vantajoso do que outro devido às características da sua implementação.

Levando-se em conta a execução na nuvem, por exemplo, um *workflow* pode ter desempenho e resultados variados de acordo com o número de máquinas virtuais instanciadas, características das imagens utilizadas, etc. Ou seja, existem diversos aspectos que podem influenciar a execução do *workflow*. Cada alternativa para uma determinada atividade pode ter um comportamento diferente de acordo com estes aspectos. Sem contar que estes aspectos podem variar durante a execução do *workflow*.

Deste cenário, surge a necessidade de um apoio dinâmico para a escolha de uma atividade alternativa. Ou seja, a possibilidade de adaptação, alterando as atividades de um *workflow* em execução sem que se faça necessário interromper a execução e reexecutar o mesmo por completo para que as alterações sejam consideradas. A necessidade de se trabalhar com *workflows* dinâmicos foi levantada em (Gil *et al.* 2007) como um desafio, porém até hoje esse problema se mantém aberto.

Com as soluções existentes hoje, os cientistas precisam interromper ou esperar o término da execução para realizar ajustes no *workflow* para então re-executar o mesmo por completo. Realizamos um levantamento recente sobre o estado da arte em monitoramento e adaptação em sistemas de *workflows* científicos onde foi constado o apoio incipiente ao monitoramento e adaptação por parte do cientista em próxima interação com a execução do *workflow* que demanda dias ou meses em execução (Mattoso *et al.* 2013). Esta adaptação em tempo de execução está diretamente ligada à produtividade da equipe que acompanha o experimento científico, conforme medições em Dias (2013).

Cada execução do *workflow* pode demandar um tempo elevado de execução. Atrelado ao tempo de execução pode existir ainda o custo financeiro, pois o tempo utilizado para execução de um *workflow* em uma nuvem computacional, por exemplo, é cobrado sob demanda. Neste cenário, um tratamento dinâmico para os *workflows* científicos pode trazer economia permitindo que os cientistas realizem ajustes no *workflow* e poupem a execução de parte do *workflow* novamente. Um dos objetivos deste trabalho é permitir que os ajustes realizados se reflitam em tempo real sobre os *workflows* em execução no momento, com o intuito de apoiar o cientista a obter os resultados finais do experimento mais rapidamente e com maior qualidade.

4. PROPOSTA DE MONITORAMENTO E ADAPTAÇÃO EM *WORKFLOWS* CIENTÍFICOS

Sistemas de *workflows* científicos já mostraram sua generalidade e tendência a padrões. Entretanto, na área de exploração e produção de petróleo o uso de sistemas de *workflows* científicos é muito incipiente. Um dos motivos está no fato de que a área de petróleo tem larga tradição de uso de computação paralela de alto desempenho e muitos sistemas de *workflows* não oferecem paralelismo em vários níveis, ou seja, os sistemas de *workflows* paralelos provêm a paralelização de dados sobre códigos sequenciais. Os sistemas de *workflows* atuais não permitem que uma atividade do *workflow* já seja um código que requer uma execução paralela.

Essa proposta visa à discussão de desafios em soluções que apoiem a gerência desse fluxo de dados, oferecendo mecanismos de monitoramento, consulta aos dados científicos e interferência na execução de experimentos científicos modelados como *workflows*.

Mesmo com toda a evolução obtida com os sistemas de *workflows* científicos, o apoio ao chamado *workflow* dinâmico, com intervenção do usuário é um dos desafios que permanece em aberto, dentre os levantados em (Gil *et al.* 2007), conforme identificamos em (Mattoso *et al.* 2013). A interação entre o usuário e o *workflow* em ambientes de alto desempenho está associada a três questões principais envolvendo a gerência do *workflow*: monitoramento da execução, análise dos resultados parciais e interferência dinâmica na execução (Mattoso *et al.* 2013). Sob esta perspectiva, a abordagem defendida nesse artigo visa a contribuir na direção da obtenção de um ambiente pleno de *workflows* dinâmicos.

Começamos a investir em colocar o cientista no processo iterativo e interativo ao longo dos últimos quatro anos sobre os sistemas de *workflows* que desenvolvemos em código aberto, a saber, o Chiron (Ogasawara *et al.* 2011) e o SciCumulus (Oliveira *et al.* 2010), com capacidade de processamento paralelo em máquinas de alto desempenho (Ogasawara *et al.* 2013) e o paralelismo com elasticidade de recursos em nuvens computacionais (Oliveira *et al.* 2012) e recuperação de falhas (Costa *et al.* 2012). Ambos os sistemas seguem uma abordagem algébrica para a representação do *workflow*, fazendo assim com que as máquinas sejam orientadas aos dados representados como tuplas em um banco de dados (Ogasawara *et al.* 2011), o que vem facilitando as contribuições obtidas nesses aspectos de *workflows* dinâmicos. Um ambiente de execução, denominado AWARD (Assunção *et al.* 2012) também se propõe a usar tuplas para gerenciar o *workflow* científico, prover o monitoramento ao usuário e algum apoio a mudanças na especificação do *workflow*. Entretanto, AWARD não disponibiliza dados de proveniência e sua abordagem fica fora do padrão PROV, dificultando a interoperabilidade entre análises de dados gerados em diferentes sistemas de *workflows*.

Ao usarmos a execução orientada a tuplas em conjunto com os dados de proveniência, é possível monitorar a execução (Pintas *et al.* 2012) por meio de consultas a dados de proveniência (Costa *et al.* 2013), fazer a visualização analítica de resultados parciais (Horta *et al.* 2013) e eventualmente modificar configurações por meio de operações sobre as tuplas (Santos *et al.* 2013).

Tais resultados foram explorados tanto nas aplicações de petróleo quanto nas aplicações de bioinformática e em mineração de textos sobre metadados de bibliografias (Dias *et al.* 2013, Mattoso *et al.* 2014), evidenciando a adequação das soluções e a generalidade das mesmas. Resultados mais específicos sobre essas aplicações podem

ser obtidos em (Bareiro *et al.* 2014, Chirigati *et al.* 2012, Dias *et al.* 2011, Dias *et al.* 2013, Guerra *et al.* 2012, Horta *et al.* 2013, Mattoso *et al.* 2014, Ocaña *et al.* 2011, 2012, 2013, Oliveira *et al.* 2011, 2012, 2013, Santos *et al.* 2013).

As experiências obtidas evidenciaram também muitos problemas em aberto, conforme apresentado em Dias (2013) e discutido a seguir. Para especificar um *workflow* científico que possa usufruir de capacidades dinâmicas, devemos possibilitar a concepção de experimentos na forma de um *workflow* iterativo com operadores algébricos específicos para avaliação de resultados intermediários, tarefa fundamental em experimentos de petróleo e gás. Ao permitir ciclos e mudanças nos *workflows* modelados através da álgebra de *workflows*, possibilitamos a modelagem de algoritmos iterativos e de experimentos com necessidades adaptativas ou de ajuste fino. Além disso, devem ser desenvolvidos módulos de composição e disparo de *workflows* para facilitar a implantação de *workflows* em ambientes de alto desempenho. Para a execução dos experimentos, devemos utilizar um modelo de execução dinâmico que apoie os ciclos dinâmicos com atividades de avaliação dos dados intermediários por parte do cientista.

Consideramos que essa participação ativa do cientista ao longo de execuções demoradas e complexas possui desafios que envolvem diversas disciplinas da computação. Tais desafios possuem diversas semelhanças com os desafios apresentados na gerência e análise de *Big Data*, por Jagadish *et al.* (2014), entre eles a participação do usuário no processo analítico. De acordo com Jagadish *et al.* (2014), os desafios na análise de dados em larga escala estão relacionados à heterogeneidade de dados; escala- grande volume de dados demandando processamento paralelo, incluindo computação em nuvens; inconsistência- se os dados não estão relacionados são necessários métodos para a descoberta de relacionamentos implícitos; tempo real - dados precisam ser filtrados e sumarizados durante a sua geração em tempo de execução, adaptando o que foi definido originalmente; privacidade e posse de dados; perspectiva do ser humano- para que o potencial de análise de *Big Data* seja obtido é necessário que a escala seja considerada não apenas sob a perspectiva computacional, mas também de seres humanos.

Ainda de acordo com Jagadish *et al.* (2014), *"In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding"*. Nesse sentido, eles sugerem que a análise de *Big Data* não seja feita totalmente de forma automática, mas que inclua o ser humano no processo (*HIL*), considerando que essa participação deva ser oferecida em todos os estágios do ciclo de vida do processo analítico em *Big Data*. Embora *Big Data* esteja mais associado às áreas de aplicação que não são as clássicas científicas, essas questões estão presentes no que estamos chamando de três questões analíticas envolvendo o ciclo de vida do *workflow* científico: monitoramento da execução, análise dos resultados parciais em tempo real e interferência

dinâmica na execução do *workflow*. Soluções nessas direções envolve uma pesquisa multidisciplinar entre áreas da computação como bancos de dados, linguagens de programação, algoritmos distribuídos, processamento paralelo, Interação humano-computador, computação gráfica dentre outras. Abordaremos algumas dessas questões na próxima seção.

5. PLANO PRELIMINAR DE SOLUÇÕES

Desafios em computação para implementar a abordagem de monitoramento e adaptação na transformação dos dados, abordada neste artigo, envolve pesquisas nas diversas áreas da computação. A seguir apresentamos alguns tópicos importantes associados a cada área em que a pesquisa deve focar nos próximos anos.

1. Engenharia de *Software*

   - Pesquisas no processo de desenvolvimento de *software* utilizado modelagem de *workflows* científicos.

   - Pesquisas sobre gerência de configuração para registrar a evolução da especificação do *workflow* e dos dados de proveniência.

2. Banco de Dados

   - Pesquisa na gerência de fluxo de dados científicos representados como tuplas.

   - Pesquisa sobre métodos de consulta a dados científicos.

   - Pesquisa sobre indexação para acesso eficiente a partes específicas dos dados científicos.

   - Pesquisa sobre o uso eficiente de bancos de dados de grafos para armazenamento e consultas a grandes conjuntos de dados de proveniência.

3. Linguagens de Programação

   - Pesquisa sobre modelos de linguagens de programação que facilitem a otimização da execução de *workflows* em tempo de execução.

   - Pesquisa sobre representação de *workflows* em vários níveis de abstração e mapeamentos corretos entre os níveis.

4. Interação Humano-Computador

   - Pesquisas em interfaces para de apoio ao HIL.

   - Pesquisas sobre formas de interação com *software* em execução, para monitoramento e adaptação.

5. Computação Paralela e de Alto Desempenho

- Pesquisa sobre o escalonamento adaptativo de tarefas com fluxo de dados/ tarefas e proveniência.

- Pesquisa sobre a execução de *workflows* em ambientes de processamento de alto desempenho reagindo à interferência do usuário.

6. Otimização

- Pesquisa em uso de funções multi-objetivo para escalonamento de recursos.

- Pesquisa em métodos de otimização baseados em meta-heurísticas para dimensionamento de recursos em ambientes de alto desempenho.

7. Computação Autonômica

- Auxilio à configuração e adaptação de *workflows*.

8. Computação Gráfica

- Pesquisa em visualização parcial do fluxo de dados.

- Pesquisa sobre métodos eficientes para comparação de visualizações e resultados intermediários.

REFERÊNCIAS BIBLIOGRÁFICAS

Adams, B.M., Bohnhoff, W.J., Dalbey, K.R., Eddy, J.P., Eldred, M.S., Gay, D.M., Haskell, K., Hough, P.D., and Swiler, L.P. (2010) "DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.0 User's Manual," Sandia Technical Report SAND2010-2183

Assuncao, L. Goncalves, C. Cunha, J.C. (2012), Autonomic Activities in the Execution of Scientific Workflows: Evaluation of the AWARD Framework, In: *Proceedings of 9th International Conference on Autonomic & Trusted Computing (UIC/ATC)*, p. 423 – 430

Bareiro, S. B. ; Ocana, K. A. C. S. ; Oliveira, Daniel ; Dias, J. ; Mattoso, Marta (2014). Exploring Large Scale Receptor-Ligand Pairs in Molecular Docking *Workflows* in HPC Clouds. In: 1*3th IEEE International Workshop on High Performance Computational Biology*, p. 536-545.

Chirigati, Fernando, Silva, Vítor, Ogasawara, Eduardo, De Oliveira, Daniel, Dias, Jonas, Porto, Fábio, Valduriez, Patrick, Mattoso, Marta (2012). Evaluating parameter sweep *workflows* in high performance computing In: *1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies - SWEET '12*. New York: ACM Press. p.1 -

Costa, F., Oliveira, D., Ocaña, K., Ogasawara, E. and Mattoso, M. (2012). Enabling Re-Executions of Parallel Scientific *Workflows* Using Runtime Provenance Data. In: 4th International Provenance and Annotation Workshop.

Costa, F., Silva, V., De Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J. and Mattoso, M. (2013). Capturing and Querying *Workflow* Runtime Provenance with PROV: A Practical Approach. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. , EDBT '13. ACM.

Coutinho, R., Drummond, L., Frota, Y., Oliveira, D., Ocaña, K., (2014), "Evaluating Grasp-based Cloud Dimensioning for Comparative Genomics: a Practical Approach". In: Proc. of the Second International Workshop on Parallelism in Bioinformatics Second International Workshop on Parallelism in Bioinformatics, Madrid, Spain.

Deelman, E., Gannon, D., Shields, M., Taylor, I., (2009), "Workflows and e-Science: An overview of workflow system features and capabilities", Future Generation Computer Systems, v. 25, n. 5, p. 528 – 540.

Deelman, E., Mehta, G., Singh, G., Su, M.-H., Vahi, K., (2007), "Pegasus: Mapping Large-Scale Workflows to Distributed Resources", Workflows for e-Science, Springer, p. 376–394.

Dias, J., Ogasawara, E., Oliveira, D., Porto, F., Coutinho, A. and Mattoso, M. (2011). Supporting Dynamic Parameter Sweep in Adaptive and User-Steered *Workflow*. In *6th Workshop on Workflows in Support of Large-Scale Science*. , WORKS '11. ACM.

Dias, J., Ogasawara, E., Oliveira, D., Porto, F., Valduriez, P. and Mattoso, M. (2013). Algebraic Dataflows for Big Data Analysis. In *Proceedings of the IEEE International Conference on Big Data*. p. 150-155.

Dias, J., (2013) Execução Interativa de Experimentos Científicos Computacionais em Larga Escala, Tese de doutorado, PESC-COPPE-UFRJ.

Fahringer, T., Prodan, R., Rubing Duan, Nerieri, F., Podlipnig, S., Jun Qin, Siddiqui, M., Hong-Linh Truong, Villazon, A., et al., (2005), "ASKALON: a Grid application development and computing environment". In: 6th IEEE/ACM International Workshop on Grid Computing, p. 122–131, Seattle, Washington, USA.

Freire, J., Koop, D., Santos, E., Silva, C. T., (2008), "Provenance for Computational Tasks: A Survey", Computing in Science and Engineering, v.10, n. 3, p. 11–21.

Gannon, D., Plale, B., Marru, S., Kandaswamy, G., Simmhan, Y. and Shirasuna, S. (2007). Dynamic, Adaptive *Workflows* for Mesoscale Meteorology. *Workflows for e-Science*. Springer. p. 126–142.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., et al., (2007), "Examining the Challenges of Scientific Workflows", Computer, v. 40, n. 12, p. 24–32.

Guerra, G., Rochinha, F. A., Elias, R., De Oliveira, D., Ogasawara, E., Dias, J. F., Mattoso, M. and Coutinho, A. L. G. A. (2012). Uncertainty Quantification in Computational Predictive Models for Fluid Dynamics Using *Workflow* Management Engine. *International Journal for Uncertainty Quantification*, v. 2, n. 1, p. 53–71.

Hanzich, M., Rodriguez, J. E., Gutierrez, N., Puente, J., Cela, J. M., (2014), "Using HPC Software Frameworks for Developing BSIT: A Geophysical Imaging Tool". In: Proceedings of the 11th World Congress on Computational Mechanics, p. 181–189

Horta, F., Dias, J., Elias, R., Oliveira, D., Coutinho, A. L. G. A., Mattoso, M., (2013), "Prov-Vis: Large-Scale Scientific Data Visualization Using Provenance (Abstract)". In: Proceedings of the International Conference on High Performance Computing, Networking, Storage and AnalysisHigh Performance Computing, Networking, Storage and Analysis, 2013 SC Companion:, Denver, CO, USA.

Jagadish, H. V., Gehrke, J.; Labrinidis, A.; Papakonstantinou, Y.; Patel, J.; Ramakrishnan, R.; Shahabi, C; (2014) "Big data and its technical challenges". *Commun. ACM* 57(7): 86-94

Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Ogasawara, E., Oliveira, D., Cruz, S. M., Martinho, W., Murta, L., (2010), "Towards supporting the life cycle of large scale scientific experiments", International Journal of Business Process Integration and Management, v. 5, n. 1, p. 79.

Mattoso, M., Ocaña, K., Horta, F., Dias, J., Ogasawara, E., Silva, V., Oliveira, D., Costa, F., Araújo, I. (2013) "User-steering of HPC *workflows*" In: *2nd ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies - SWEET '13*. New York: ACM Press. p.1-4

Mattoso, M ; Dias, J. ; Costa, F. ; Oliveira, D. ; Ogasawara, E. (2014). Experiences in using provenance to optimize the parallel execution of scientific *workflows* steered by users. In: *Provenance Analytics Workshop*, Colônia. Provenance Week.p. 1-4

Moreau, L., Missier, P., Belhajjame, K., Cresswell, S., Golden, R., Groth, P., Miles, S., Sahoo, S., (2011). The PROV Data Model and Abstract Syntax Notation. Disponível em: http://www.w3.org/TR/prov-dm/. Acesso em: 14 Dec 2011.

H. Nguyen and D. Abramson, 2012, WorkWays: Interactive workflow-based science gateways, In: *Proceedings of the 8th IEEE International Conference on E-Science (e-Science)*, p. 1–8

Ocaña, K. A. C. S., Oliveira, D., Dias, J., Ogasawara, E. and Mattoso, M. (7 dec 2011a). Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific *Workflow*. In *2011 IEEE Seventh International Conference on e-Science (e-Science)*. Estocolmo, Suécia. IEEE.

Ocaña, K. A. C. S., Oliveira, D., Ogasawara, E., Dávila, A. M. R., Lima, A. A. B. and Mattoso, M. (2011b). SciPhy: A Cloud-Based *Workflow* for Phylogenetic Analysis of Drug Targets in Protozoan Genomes. In: Norberto de Souza, O.; Telles, G. P.; Palakal, M.(Eds.). *Advances in Bioinformatics and Computational Biology*. Berlin, Heidelberg: Springer. v. 6832p. 66–70.

Ocaña, K. A. C. S., Oliveira, D. De, Horta, F., Dias, J., Ogasawara, E. and Mattoso, M. (2012a). Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific *Workflow*. *Advances in Bioinformatics and Computational Biology*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. v. 7409p. 179–191.

Ocana, Kary A. C. S., De Oliveira, Daniel, Dias, Jonas, Ogasawara, Eduardo, Mattoso,Marta (2012b). Discovering drug targets for neglected diseases using a pharmacophylogenomic cloud *workflow* In: *IEEE 8th International Conference on EScience (eScience)*, Chicago, Estados Unidos. p.1 - 8

Ocaña, K. A. C. S., Oliveira, F., Dias, J., Ogasawara, E. and Mattoso, M. (2013). Designing a parallel cloud based comparative genomics *workflow* to improve phylogenetic analyses. *Future Generation Computer Systems*, v. 29, n. 8, p. 2205–2219.

Ogasawara, E., Dias, J., Oliveira, D., Porto, F., Valduriez, P. and Mattoso, M. (2011). An Algebraic Approach for Data-Centric Scientific *Workflows*. *Proc. of VLDB Endowment*, v. 4, n. 12, p. 1328–1339.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., Oliveira, D., Porto, F., Valduriez, P. and Mattoso, M. (2013). Chiron: A Parallel Engine for Algebraic Scientific *Workflows*. *Concurrency and Computation*, v. 25, n. 16, p. 2327–2341.

Oliveira, D. ; Ogasawara, E.; Baião, F. ; Mattoso, M., (2010a). SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific *Workflows*. In: CLOUD 2010, 2010, Miami, Estados Unidos. p. 378-385.

Oliveira, D. Baião, F. ; Mattoso, M., (2010b). Towards a Taxonomy for Cloud Computing from an e-Science Perspective. In: Nikolaos Antonopoulos; Lee Gillam. (Org.). Cloud Computing: Principles, Systems and Applications. : Springer, 2010, v. 3, p. 47-62.

Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J., Baião, F. and Mattoso, M. (2011). A Performance Evaluation of X-Ray Crystallography Scientific *Workflow* Using SciCumulus. In *IEEE International Conference on Cloud Computing (CLOUD)*, Washington, Estados Unidos.

Oliveira, D., Ocaña, K., Baião, F. and Mattoso, M. (2012). A Provenance-based Adaptive Scheduling Heuristic for Parallel Scientific *Workflows* in Clouds. *Journal of Grid Computing*, v. 10, n. 3, p. 521–552.

Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J., Goncalves, J., Mattoso, Marta (2012). Cloud based Phylogenomic Inference of Evolutionary Relationships: A Performance Study In: *International Workshop on Cloud Computing and Scientific Applications*, **Best Paper Award**, Ottawa, Canadá. CCGrid. IEEE/ACM

Oliveira, D., Ocaña, K. A. C. S., Ogasawara, E., Dias, J., Gonçalves, J., Baião, F. and Mattoso, M. (2013). Performance evaluation of parallel strategies in public clouds: A study with phylogenomic *workflows*. *Future Generation Computer Systems*, v. 29, n. 7, p. 1816–1825.

Pintas, J., Oliveira, D., Ocaña, K., Dias, J., Mattoso, M., (2012), "Monitoramento em Tempo Real de Workflows Científicos Executados em Paralelo em Ambientes Distribuídos". In: VI e-Science workshopXXXII Congresso da Sociedade Brasileira da Computação, Curitiba, Paraná, Brazil.

Santos, I., Dias, J., Oliveira, D., Ogasawara, E., Ocaña, K., Mattoso, M., (2013), "Runtime Dynamic Structural Changes of Scientific Workflows in Clouds". In: Proceedings of the IEEE/ACM 6th International Workshop on Clouds and (eScience) Applications Management - CloudAM, p. 417–422, Dresden, Alemanha.

SBC (2006), Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016, Relatório sobre o Seminário dos Grandes Desafios da Computação, São Paulo, Brasil 2006

Taylor, I., Shields, M., Wang, I., Harrison, A., (2007), "The Triana Workflow Environment: Architecture and Applications", Workflows for e-Science, Springer, p. 320–339.

Wozniak, J. M., Armstrong, T. G., Wilde, M., Katz, D. S., Lusk, E., Foster, I. T., (2013), "Swift/T: Large-Scale Application Composition via Distributed-Memory Dataflow Processing". In: Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), p. 95–102

DESENVOLVIMENTO DE SISTEMAS DE SOFTWARE PARA AUMENTO DA SEGURANÇA NA CADEIA DE MINERAÇÃO

Cleidson R. B. de Souza[1, 2], Schubert Carvalho[1], Gustavo Pessin[1]

1. INTRODUÇÃO

Este trabalho apresenta uma proposta visando participação no "3º Seminário dos Grandes Desafios em Computação – Fase 2 – Ênfase em Grandes Desafios do Mercado e do Governo" visando a identificação de possíveis parcerias entre a Vale S.A. através do Instituto Tecnológico Vale e outros membros da academia, do governo e da indústria, no contexto de "Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos para Aumento da Segurança[1] na Cadeia da Mineração", ou simplesmente, "Desenvolvimento de Sistemas de Software para Aumento da Segurança na Cadeia de Mineração".

Uma das principais motivações desta proposta é o trabalho de Câmara (2012) que indica que as áreas de maior produção científica brasileira são aquelas associadas à "economia dos recursos naturais". Entretanto, que a ciência da computação não está listada entre estas áreas. Isto sugere que uma das formas da Ciência da Computação brasileira aumentar sua produção científica inclui o direcionamento, pelo menos parte, de seus esforços para a agenda interna brasileira, em especial aspectos relacionados aos recursos naturais como a agricultura, bioinformática, mineração entre outros. Além disso, é importante ressaltar que a segurança dos empregados e das comunidades é um aspecto de fundamental importância para a Vale S.A. que investe significativamente neste aspecto. De fato, um dos cinco valores fundamentais da mesma é a "A Vida em Primeiro Lugar". Mais do que isso, segurança é também um fator essencial em outras atividades econômicas relevantes para o Brasil como por exemplo, nas áreas de petróleo e energia. Isto demonstra o potencial que tecnologias desenvolvidas para aumento da segurança em uma área possam ser utilizadas em outras atividades econômicas importantes para o país.

O restante deste texto está organizado da seguinte forma. A próxima seção descreve a motivação para esta proposta e o contexto no qual a mesma está inserida. Depois

[1]Instituto Tecnológico Vale
Rua Boaventura da Silva, 955, Belém, PA – Brasil
[2]Faculdade de Computação, Universidade Federal do Pará
Rua Augusto Côrrea, 1, Belém, PA – Brasil

cleidson.desouza@acm.org, schubert.carvalho@itv.org, gustavo.pessin@itv.org

[1]Neste caso, estamos falando em segurança como safety e não como security.

disso, a seção 3 descreve o desafio proposto pelos autores de um modo geral e também através de exemplos específicos da cadeia de mineração. Finalmente, a seção 4 apresenta as conclusões bem como o resultado esperado desta proposta.

2. MOTIVAÇÃO E CONTEXTO

A Vale S.A., assim como qualquer outra empresa de liderança no mercado, necessita de investimentos em pesquisas que vislumbrem tendências e se antecipem a problemas que futuramente possam afetar os seus negócios. Pesquisas que resultem em uma visão de longo prazo garantem a dianteira no desenvolvimento tecnológico e um melhor aproveitamento de seus investimentos resultando em vantagem competitiva diante dos desafios e oportunidades. Diante desse cenário, a Vale investe continuamente em pesquisa e inovação através de parcerias institucionais bem como através de seus próprios centros de pesquisa, dentro deles o Instituto Tecnológico Vale (ITV), concebido pela Vale com o objetivo de produzir pesquisas científicas e tecnológicas que tenham potencial inédito e transformador.

Uma outra maneira de garantir a vantagem competitiva das empresas é por meio da utilização de tecnologias de informação e comunicação (também simplesmente chamadas de *TIC*), especialmente sistemas de software. De fato, os modernos sistemas de software tornaram-se vitais para os negócios de todos os tipos de organizações e alguns autores argumentam que eles são fundamentais para a humanidade como um todo[2]. O aumento do uso de sistemas de software em contextos diferentes e cada vez mais abrangentes torna a construção de software uma atividade cada vez mais importante, e também mais complexa visto que diversos serviços dependem destes sistemas de software.

No contexto da mineração, sistemas de software também podem ser considerados essenciais conforme sugerido pelo programa TI Maior[3] que reconhece o "Mercado de Software para Mineração" como um dos seus ecossistemas digitais. Ainda no contexto do programa TI Maior, outros ecossistemas também podem ser relacionados com as atividades da cadeia de mineração como:

O "Mercado de Software para Tecnologias Estratégicas", no contexto de internet das coisas através do desenvolvimento de soluções de monitoramento de ativos ou de segurança para os empregados envolvidos em atividades de mineração;

O "Mercado de Software para Energia" é de grande relevância dada a demanda energética para transformar recursos minerais em prosperidade e desenvolvimento. O caminho

[2]De acordo com Bjarne Stroustrup: "Our civilization runs on software".
[3]Ver http://timaior.mcti.gov.br

de modernização das redes atuais rumo aos "*Smart Grids*" (redes inteligentes) nada mais é do que a aplicação direta das tecnologias de software e comunicação às redes atuais (Singhal e Saxena, 2012). Dada a escala das operações da Vale S.A., qualquer ganho energético marginal possui um potencial multiplicador imenso; e

Finalmente, o "Mercado de Software para Agricultura e Meio Ambiente" no qual tecnologias de bioinformática podem ser usadas no auxílio as atividades de licenciamento ambiental, recuperação de áreas degradadas, monitoramento ambiental entre outros.

A Austrália pode ser usada como um exemplo claro da importância dos sistemas de software nas atividades de mineração (Wikipedia, 2013):

"Australia's high labour costs and first-world safety regulations, distinctive geology, and the importance placed on mining research by successive governments and businesses has meant that the Australian mining sector is quite technologically advanced. A large proportion of mines worldwide make use of Australian-developed computer software, such as specialised Geological Database and Resource Estimation Modelling software by Micromine and geology/mine planning software by Runge Ltd and Maptek Pty Ltd.

Australia is also home to promising new tech companies that offer mine planning software including Oreology and Paradyn. Mines in Australia are leading the market globally deploying mine production data management software such as Corvus developed by Intov8 Pty Ltd, which displays real-time production data from multiple source systems on dashboards, and includes comprehensive dynamic analysis and reporting, driving process and cost efficiencies at the shift level. Australia's mining services, equipment, and technology exports are over $2 billion annually."

É importante ressaltar a última frase do texto acima que destaca o valor das exportações da Austrália em serviços, tecnologias e equipamentos, assim ilustrando o potencial econômico dos sistemas de software e de outras tecnologias e serviços voltados para a cadeia de mineração. De maneira similar, a Figura 1 ilustra o trabalho de Câmara (2012) que sugere que as áreas de maior produção científica brasileira são aquelas associados à "economia dos recursos naturais"[4]. Em outras palavras, além do potencial impacto na economia, a ciência da computação brasileira pode também se beneficiar de uma maior produção cientifica se direcionar esforços para aspectos relacionados aos recursos naturais.

A Vale S.A., bem como o ITV, reconhecem a importância estratégica dos sistemas de software e, de um modo mais geral, das tecnologias de informação e comunicação como mecanismos para aumento de competitividade e produtividade, redução de custos entre

[4]É importante também observar na Figura 1 que a produção científica brasileira da área de Ciência da Computação está mais associada a agendas externas de pesquisa do que uma agenda de pesquisa propriamente brasileira.

outros benefícios. De fato, o ITV tem diversos pesquisadores cuja principal área de atuação são as TICs incluindo áreas como engenharia de software, redes de sensores, aprendizado de máquina, robótica, telecomunicações, otimização, sistemas de informação geográfica e processamento de imagens. Estes pesquisadores atuam em diversos projetos de pesquisa financiados pela Vale S.A. em cooperação com universidades e empresas do Brasil e exterior, bem como com financiamento do CNPq através do Edital MCTI/CT-Info/CNPq n. 59/2013.

3. O DESENVOLVIMENTO DE SISTEMAS DE SOFTWARE PARA AUMENTO DA SEGURANÇA NA CADEIA DE MINERAÇÃO

Neste trabalho, a cadeia de mineração pode ser entendida de maneira simplificada como as 3 grandes atividades realizadas pela Vale S.A., a saber: extração, transporte e armazenamento de minérios[5]. Ou ainda, no caso da Vale pode-se falar em 2 sítios diferentes onde estas atividades ocorrem: Minas (extração e armazenamento) e Portos (armazenamento). Além disso, o transporte de minérios é feito por meio de ferrovias (das minas para os portos) ou por meio de navios (dos portos para seus clientes).

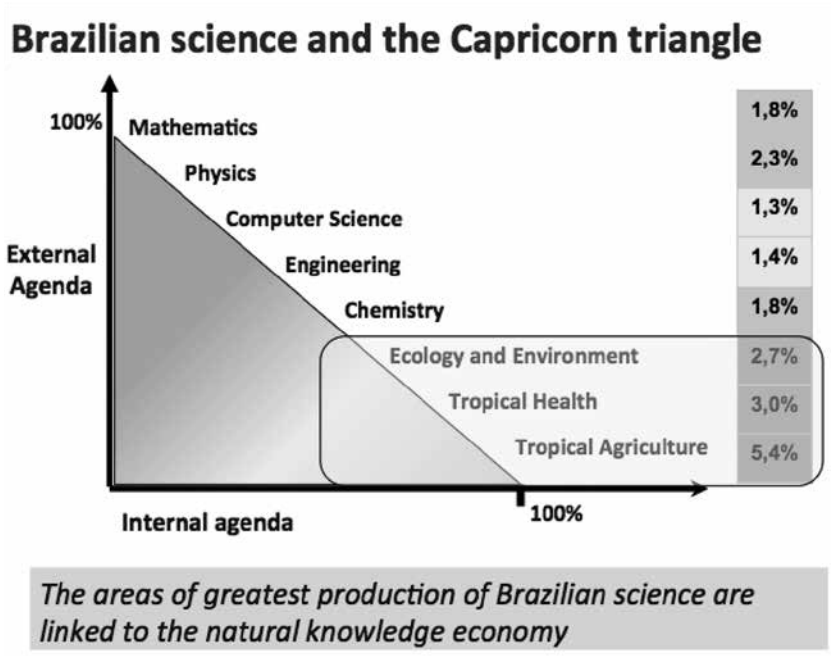


Figura 1 - A ciência brasileira e o triângulo de Capricórnio (Câmara, 2012)

[5]Obviamente, outras empresas de mineração possuem suas peculiaridades, mas no contexto deste trabalho consideraremos esta descrição geral.

Diversas atividades da cadeia de mineração apresentam riscos tanto para os empregados da Vale e das empresas terceirizadas quanto para as pessoas que vivem ao redor de suas instalações. Além disso, algumas atividades também podem colocar em risco tanto o meio ambiente (por meio de contaminações) como os ativos da empresa (danos aos equipamentos). Desta forma, a Vale investe continuamente para aumentar a segurança de suas atividades através da melhoria de seus processos, treinamento de seus empregados, desenvolvimento e aquisição de tecnologias (inclusive de TICs) entre outras formas. É importante ressaltar que a segurança dos empregados e das comunidades é um fator fundamental em outras atividades da economia de recursos naturais como, por exemplo, na área de Petróleo e Energia, isto é, existe uma possibilidade concreta que as tecnologias desenvolvidas no contexto das atividades da cadeia de mineração possam ser utilizadas em outras atividades econômicas.

Entretanto, em relação ao desenvolvimento de tecnologias de TICs para aumento da segurança, existem vários desafios que precisam ser investigados, por isto a necessidade de identificação de parceiros acadêmicos, industriais e governamentais. Assim, o restante desta seção visa ilustrar alguns destes desafios.

3.1. O Desenvolvimento de Sistemas de Software Resilientes e de Alta Confiabilidade para a Internet das Coisas

Devido a redução no custo dos sensores e ao aumento da capacidade de processamento, é possível observar nos últimos anos o surgimento de tecnologias computacionais cada vez menores, interconectadas e ubíquas. Estas tecnologias visam a conexão não apenas de sistemas computacionais, mas também de itens mundanos do dia a dia das pessoas, dos carros e das cidades, criando assim a chamada internet das coisas. Atualmente, diversas universidades, empresas e institutos de pesquisa estão desenvolvendo estas tecnologias e explorando a aplicação das mesmas em diferentes contextos. A GE Software, por exemplo, têm utilizado a internet das coisas como um mecanismo para coletar e processar informações sobre os diversos equipamentos que ela produz visando o aumento da eficiência e eficácia destes equipamentos (GE Software, 2013) sendo por este motivo considerada uma das 50 empresas mais inovadoras do mundo (Fast Company, 2014).

Entretanto, o desenvolvimento destas tecnologias e sua efetiva implantação não é uma atividade trivial. O contexto tradicional de desenvolvimento de software (sistemas corporativos, Web ou mais recentemente aplicativos para dispositivos móveis) é bastante diferente do contexto no qual as tecnologias da internet das coisas irão atuar. Neste caso, os cenários possíveis envolvem dezenas de milhares de sensores com poder de processamento bastante variado e que podem enviar informações de tamanhos bastante diferentes (desde bytes até megabytes) com frequências de transmissão e com mecanismos de resiliência variados.

Em particular, no caso de sistemas de software que visam a segurança de pessoas, estes diferentes contextos tornam-se extremamente importantes, pois requerem a sua correta identificação, especificação, modelagem, implementação e validação. Por exemplo, pode-se imaginar um sensor que indica a localização de uma pessoa dentro um determinado sítio operacional. Neste caso, este sensor envia informações relativamente simples (a localização das pessoas) com uma frequência média de segundos. Em outro cenário, pode-se imaginar um sensor acoplado na traseira de um trem que possui uma câmera de vídeo e transmite em tempo real megabytes de informação a serem processados para identificar possíveis obstáculos para o trem. Este mesmo sensor precisa ser capaz de receber o resultado do processamento destas informações para assim realizar as ações adequadas, que podem ser entre outras, informar ao maquinista através de interfaces gráficas adequadas, reduzir a velocidade do trem de maneira automática ou semi-automática, etc. *Assim, um dos desafios aqui colocados é a identificação, especificação e modelagem de requisitos em sistemas de software no contexto da internet das coisas*. Além disso, deve-se ter em mente que tais atividades precisam estar associadas a mecanismos de tolerância a falhas que permitam que tais sistemas sejam resilientes. Entretanto, surge a pergunta: *como fornecer mecanismos de tolerância a falhas que possam ser adequados a sistemas de software que atuam em contextos tão diferentes?* Alguns autores começam a indicar possíveis soluções para isto, como por exemplo (Friedrich 2010; Beder, 2013). *Um terceiro desafio ainda neste contexto é como fazer a integração dos dados de diferentes sensores de tal forma que estes sistemas possam ser considerados confiáveis, isto é, com a garantia de entrega de dados em um determinado intervalo de tempo*.

Neste exemplo, o primeiro autor deste trabalho têm adotado em suas pesquisas o método de pesquisa da etnografia, que é o principal método que a antropologia utiliza para coletar e analisar dados (McGrath, 1998). A etnografia proporciona um entendimento acerca das reais atividades desempenhadas por um grupo de atores ao invés das atividades formais, normalmente relatadas em entrevistas, ou mesmo as atividades prescritas em um manual ou processo (Blomberg et al, 1993). Em razão disto, a etnografia tem sido adotada por profissionais da área de IHC e CSCW há décadas (ver Suchman (1987)) visando caracterizar o ambiente de trabalho das pessoas e, assim, identificar os problemas enfrentados por estas pessoas em seu dia a dia. Desta forma, a etnografia poderá ser usada para entender exatamente as atividades realizadas pelos diferentes atores envolvidos na da cadeia de mineração. Isto auxiliará na especificação do contexto dos mesmos, bem como no *design* das tecnologias para os mesmos (Blomberg et al, 1993).

3.2. Desenvolvimento e Definição de Políticas de Utilização de Grupos de Veículos Autônomos

A dinâmica indústria de mineração não é uma exceção às outras indústrias e a adoção gradual de soluções inovadoras de automação representa um caminho sem volta na

direção das chamadas minas inteligentes (*smart mines*). Embora uma infraestrutura moderna de comunicações e um alto grau de automação representem pontos fundamentais para minas inteligentes, acredita-se que a presença humana continue integrada em partes da operação, seja para controle ou para monitoramento e verificação de atividades a serem otimizadas.

Este desafio apresenta um amplo potencial para aumento de competitividade tanto para a Vale quanto para o setor de mineração brasileiro como um todo, além de permitir a criação de valor de longo prazo para a sociedade e para o meio ambiente. A pesquisa e o desenvolvimento de ferramentas de otimização, por exemplo, buscam aumentar a eficiência de diversos tipos de processos de tomada de decisão, desde a programação de operações e a alocação de recursos, até o planejamento de capacidade e o projeto de cadeias produtivas minerais. Da mesma forma, o desenvolvimento e a aplicação de técnicas e ferramentas de automação e robótica também podem trazer resultados significativos diretos na redução de custos operacionais, no aumento de produtividade e na redução da exposição ao risco dos empregados. Além disso, estas tecnologias podem apoiar o monitoramento ambiental, a avaliação de áreas de risco e a proteção de ativos de biodiversidade.

Como o próprio nome indica, a linha de pesquisa em Veículos Autônomos pressupõe o desenvolvimento de modelos matemáticos, algoritmos, heurísticas, sistemas de controle e protocolos de comunicação especializados para as diferentes aplicações a que se destinam. Nesse sentido, acredita-se que as tecnologias a serem desenvolvidas têm forte potencial de se reverterem em novos negócios de base tecnológica, favorecendo o empreendedorismo, a produção de conhecimento e a geração de renda por meio de produtos e serviços de alto valor agregado.

3.3. Redes de Sensores e Atuadores sem fio Para Monitoramento e Atuação em Regiões Críticas

A criação e a manutenção de redes de sensores em regiões críticas (como ambientes insalubres ou hostis) são desafios correntes assim como a correta interpretação, filtragem e/ou fusão dos dados recolhidos para geração de conhecimento em tempo real. Estes aspectos apresentam desafios tanto em nível de serviço como em nível de confiabilidade. A criação de software com garantia de qualidade, associado com a integração de hardware que em conjunto permitem criar redes de sensores e atuadores sem fio, apresenta ainda o desafio do emprego de técnicas de aprendizado de máquina para a correta utilização dos dados coletados e possível uso da informação para a atuação autônoma ou para a geração de alertas para os empregados envolvidos no processo.

Neste contexto, a criação de arquiteturas de software flexíveis, adaptativas e escaláveis devem prover um avanço na aplicação de redes de sensores e atuadores para ambiente insalubres, o que deve aumentar a segurança dos funcionários e também a

disponibilidade dos serviços de mineração. Aspectos de segurança de sistemas e dados também serão desafios a serem abordados, dado que tem impacto na garantia de qualidade dos serviços.

3.4. Formação de Recursos Humanos Especializados: TICs e Cadeia de Mineração

Assim como em outras partes do mundo, um dos grandes problemas da indústria de TIC brasileira é a falta de profissionais qualificados. O governo brasileiro, bem como outras instituições relevantes como a SBC e a Brasscom reconhecem esta necessidade e têm proposto diversas iniciativas visando amenizar este problema.

No caso de sistemas de software para a cadeia de mineração, o problema se torna ainda mais relevante, pois são necessários profissionais de TIC que possuem, ou estejam motivados para adquirir conhecimento sobre as atividades da cadeia de mineração incluindo aí exploração mineral, logística, segurança, entre outros. Outra alternativa é treinar profissionais da cadeia de mineração na utilização e desenvolvimento de ferramentas de TIC. Ambos os casos são desafiadores e requerem um trabalho interdisciplinar envolvendo profissionais de duas áreas do conhecimento no treinamento destes profissionais, além de conhecimento e visão de mercado para a identificação de necessidades e consequentemente de oportunidades de negócio.

Neste contexto, o Governo do Estado do Pará, através da Secretaria de Ciência e Tecnologia e Inovação representada pelo Dr. Rodrigo Reis, e o Instituto Tecnológico Vale, através do primeiro autor desta proposta, têm discutido a criação de um Curso de Residência em Desenvolvimento de Software para a Cadeia de Mineração. Este curso seria um curso de Pós-Graduação em nível de Especialização a ser realizado segundo modelo dos Editais 01/2008 e 06/2010 do CNPq em parceria com universidades nacionais. O objetivo final do curso seria o desenvolvimento de sistemas de software para apoiar atividades da cadeia da mineração. Estes sistemas seriam desenvolvidos pelos alunos do curso sob orientação dos membros das instituições envolvidas. Além de treinamento em temas da cadeia de mineração e de TICs, estes alunos também receberiam treinamento na área de empreendedorismo e inovação tecnológica visando fomentar o espírito empreendedor dos mesmos.

4. CONCLUSÕES

Este texto descreveu o desafio "Desenvolvimento de Sistemas de Software para Aumento da Segurança na Cadeia de Mineração" enquanto proposta para participação no "3º Seminário dos Grandes Desafios em Computação – Fase 2 – Ênfase em Grandes Desafios do Mercado e do Governo". Este desafio foi exemplificado através de 4 itens, a saber: (i) o desenvolvimento de sistemas de software resilientes e de alta confiabilidade para a

internet das coisas, (ii) desenvolvimento de veículos autônomos, (iii) redes de sensores e atuadores sem fio para monitoramento e atuação em regiões críticas e, finalmente, (iv) a formação de recursos humanos especializados em TICs e na cadeia da mineração.

Como pode-se observar o desafio aqui proposto requer avanços em diferentes áreas da Ciência da Computação, bem como conhecimento de outras áreas do conhecimento como a Mineração, Meio Ambiente, Engenharia de Transportes, etc. Além isto, conhecimento sobre empreendedorismo e inovação é necessário para a identificação de oportunidades de negócio que permitiriam ao Brasil alavancar suas exportações em software, aumentar a sua produção científica, e ultimamente a redução de custos para as empresas brasileiras, bem como o aumento da segurança de seus empregados e da população em geral. Assim, de uma maneira geral, esta proposta visa a identificação de possíveis parcerias entre a Vale S.A. através do Instituto Tecnológico Vale e outros membros da academia, do governo e da indústria para atuação nos desafios propostos.

AGRADECIMENTOS

Os autores gostariam de agradecer ao suporte financeiro do CNPq através do edital MCTI/CT-Info/CNPq Nº 59/2013.

REFERÊNCIAS

Beder, D. M. ; Ueyama, J.; Albuquerque, J. P.; Lordello, M. FlexFT: A Generic Framework for Developing Fault-Tolerant Applications in the Sensor Web, International Journal of Distributed Sensor Networks, 2013.

Blomberg, J., Giacomi, J., Mosher, A., and Swenton-Wall, P (1993). Ethnographic Field Methods and Their Relation to Design. In: Participatory Design: Principles and Practices. Lawrence Erlbaum Associates Inc., USA, 1993.

Câmara, G. (2012) Global change and sustainable development: towards a research agenda for Brazilian Science. Workshop de Colaboração ITV-MIT, Belém, PA, Março de 2012.

Fast Company (2014). The World's Most Innovative Companies 2014. Texto de Jon Gertner.

Friedrich, G.; Fugini, M.; Mussi, E.; Pernici, B.; Tagni, G., "Exception Handling for Repair in Service-Based Processes," Software Engineering, IEEE Transactions on , vol.36, no.2, pp.198,215, March-April 2010

GE Software (2013). The Case for an Industrial Big Data Platform Laying the Groundwork for the New Industrial Age. Available at:

McGrath, J. E. (1994). Methodology Matters: Doing Research in the Behavioral and Social Sciences, 1994.

Singhal, A.; Saxena, R. P., "Software models for Smart Grid," Software Engineering for the Smart Grid (SE4SG), 2012 International Workshop on , vol., no., pp.42,45, 3-3 June 2012.

Suchman, L. Plans and situated actions: the problem of human-machine communication. Cambridge, Cambridge University Press, 1987.

Wikipedia (2013) Texto extraído de: http://en.wikipedia.org/wiki/Mining_in_Australia em 10 de Outubro de 2013.

Promoção

SBC Sociedade Brasileira de Computação

35 ANOS

Apoio

MCTI Ministério da Ciência, Tecnologia e Inovação

Brasscom

Patrocínio

EMC²

Realização

Universidade Federal do Rio de Janeiro

NCE UFRJ Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais