FAPEMIG - Fundação de Amparo à Pesquisa do Estado de Minas Gerais
Edital N º 19/2013 - PROGRAMA DE APOIO A NÚCLEOS DE EXCELÊNCIA - PRONEX

Título do Projeto	**Modelos, Algoritmos e Sistemas para Web - MASWeb**
Coordenador	Nívio Ziviani
E-mail	nivio@dcc.ufmg.br
Endereço postal	Departamento de Ciência da Computação - Universidade Federal de Minas Gerais - Av. Antônio Carlos, 6627 - 31270-010, Belo Horizonte, MG - (31) 3409.5860
Instituição Proponente	Universidade Federal de Minas Gerais (UFMG)
Intituições Parceiras	Universidade Federal de Ouro Preto(UFOP); Universidade Federal de São João Del Rei(UFSJ); Universidade Federal de Juiz de Fora(UFJF); Centro Federal de Educação Tecnológica de Minas Gerais(CEFET-MG); Pontifícia Universidade Católica de Minas Gerais(PUC-MG)

Belo Horizonte, Março 2014

Sumário

1 Sumário Executivo

A Web é o fenômeno de mídia mais importante e revolucionário desde a invenção da imprensa por Gutenberg no século XV. A grande contribuição da Web é a democratização da produção mundial de informação e conhecimento, até então monopólio de uns poucos agentes que controlavam o que poderia ser editado e distribuído, como as editoras e os jornais. Com o advento da Web, qualquer pessoa com acesso a um servidor Web pode produzir conteúdo que fica imediatamente acessível por meio de hyperlinks e das máquinas de busca. Esse fenômeno ficou mais evidente com o surgimento de novos serviços na Web tais como YouTube, Facebook, Twitter e Flickr, que tornam as tarefas de produção e publicação de conteúdo multimídia digital muito mais fácil. Isso tem levado a uma interação mais intensa entre os usuários da Web por meio de redes sociais online, um fenômeno social sobre o qual ainda se conhece muito pouco.

Para estudar os diversos fenômenos relacionados com a Web estamos propondo o projeto Modelos, Algoritmos e Sistemas para a Web. O objetivo do projeto é desenvolver modelos, algoritmos e novas tecnologias que permitam aumentar a integração da Web com a sociedade, tornando mais efetiva e mais segura a distribuição de informação, e mais eficazes e eficientes os seus serviços, de forma a proporcionar um vetor de mudanças sociais e econômicas no País. As atividades do projeto compreendem atividades relacionadas à pesquisa, à formação de recursos humanos e à transferência de conhecimento para a sociedade e para o setor empresarial.

Dada a escala exponencial de crescimento da Web, o surgimento de novas aplicações e serviços está limitado por aspectos de infraestrutura de software e hardware. Portanto, o estudo dos diversos aspectos que envolvem a Web extrapola as atividades tradicionais de geração de conteúdo e criação de novos serviços, demandando o desenvolvimento de novas tecnologias que permeiam as suas diversas camadas, bem como o entendimento da sociedade que utiliza os seus recursos. Na nossa proposta relacionada às atividades de pesquisa, partimos de uma visão unificada da Web como sendo constituída de três camadas interdependentes de redes de relacionamentos complexas e dinâmicas pelas quais a informação flui e é disseminada. A camada constituída pelas interações sociais proporcionadas pela Web cria demandas sobre a camada de serviços por meio dos quais estas interações são realizadas. A camada de serviços, por sua vez, impõe demandas adicionais sobre a camada de infraestrutura da Web. Essas três camadas de redes compreendem um conjunto de interações entre pessoas, objetos informacionais, serviços e componentes de software e hardware.

Nossa proposta de pesquisa pretende contribuir com resultados inéditos nas três camadas de redes citadas. Para tal, pretendemos trabalhar e desenvolver soluções para três grandes desafios identificados a partir dessa visão unificada da Web: (i) Identificação, caracterização e modelagem de interesses e padrões de comportamento das pessoas na Web e das redes estabelecidas entre elas; (ii) Tratamento da informação que circula pelas diversas redes da Web, considerando as atividades de coleta, extração e processamento da informação; (iii) Entrega da informação de forma satisfatória e independente de tempo e localização geográfica.

Na nossa proposta relacionada à formação de recursos humanos, pretendemos formar um número expressivo de doutores, mestres e graduados. Consideramos a formação desses profissionais e pesquisadores como um dos principais resultados da proposta, pois abrirá grandes possibilidades de desenvolvimento da área no futuro. De fato, a equipe do projeto foi formada principalmente por pesquisadores da UFMG que tiveram atuação no Instituto Nacional de Ciência e Tecnologia para a Web

(InWeb), e grande parte dos pesquisadores de outras instituições mineiras envolvidas no projeto foram formados na UFMG, no escopo das atividades de pesquisa do Instituto. Além disso, a própria equipe da UFMG evoluiu desde o InWeb, com a incorporação de novos pesquisadores à equipe. As conexões com parceiros nacionais e internacionais, por outro lado, foi ampliada ao longo dos últimos anos, e será ainda mais reforçada pelos grupos de universidades mineiras que compõem a equipe do presente projeto. As instituições participantes incluem um programa de pós-graduação em Ciência da Computação nível 7 na Capes (UFMG), dois programas nível 4 (UFOP e PUC-MG) e um programa nível 3 (UFJF), além de um grupo emergente associado a um programa de mestrado em Modelagem Matemática e Computacional (CEFET/MG), formando um grupo de pesquisadores capacitado para a formação de recursos humanos de alto nível. Com relação ao aspecto de disseminação do conhecimento, pretendemos organizar eventos sobre temas da Web com alunos de diferentes cursos de graduação, com o objetivo de divulgar os novos conhecimentos gerados e atrair novos alunos para a área de Ciência da Computação.

Na nossa proposta relacionada à transferência de conhecimento para a sociedade e para o setor empresarial, consideramos a geração de conhecimento e o domínio de tecnologia de ponta em áreas relacionadas a Web e Redes Complexas de grande importância social e econômica para o país. O desenvolvimento de competência nacional nessas áreas cria um mercado para sistemas que explorem informação valiosa que ocorre em grande volume nos sites da Web, podendo gerar um fator de desenvolvimento econômico e social e de posicionamento estratégico do país no contexto global. A exemplo de iniciativas anteriores de integrantes da equipe deste projeto (Miner Technology Group e Akwan Information Technologies - adquirida pela Google em 2005), temos dois exemplos concretos de criação de start-ups intensivas em conhecimento, a Zunnit Technologies (www.zunnit.com.br) e a Neemu Technologies (www.nhemu.com.br). Uma última iniciativa em curso é a criação de um Centro de Tecnologia para a Web (CTWeb) no Parque Tecnológico BHTEC.

Origem e Descrição do Núcleo

O grupo de pesquisadores que propõe a formação do MASWeb é pioneiro no Brasil em pesquisas relacionadas à Web, tais como gerência de dados, recuperação de informação, sistemas distribuídos em larga escala e descoberta de conhecimento. A pesquisa tem sido desenvolvida em seis laboratórios do Departamento do Ciência da Computação da UFMG, Laboratório de Bancos de Dados (LBD), Laboratório de Tratamento da Informação (LATIN), Laboratório de Análise e Modelagem de Desempenho (CAMPS), Laboratório de Computação Paralela (LCP), Laboratório de Vídeo sob Demanda (VOD) e Núcleo de Processamento Digital de Imagens (NPDI), criados ao longo das décadas de 80 e 90, e que hoje são referências nacionais e internacionais em suas respectivas áreas de atuação.

O grupo de pesquisadores é derivado do InWeb - Instituto Nacional de Ciência e Tecnologia para Web e inclui, além de membros das equipes originais da UFMG e do CEFET-MG, novos pesquisadores que hoje integram as equipes de grupos de pesquisa nucleados nos últimos cinco anos na UFOP, UFJF, UFSJ e PUC-MG como resultado das atividades do próprio Instituto. Praticamente todos os pesquisadores das instituições parceiras foram formados na UFMG e têm mantido os vínculos de pesquisa com os seus grupos de origem na UFMG.

O grupo formou na última década cerca de 200 mestres e 30 doutores em tópicos diretamente relacionados ao tema do projeto. É importante ressaltar que a UFMG tem tido papel nucleador de destaque mesmo fora do estado, tendo sido a base da formação do grupo do Instituto de Computação da UFAM, cuja pós-graduação alcançou recentemente nível 5 segundo a avaliação da CAPES. Acreditamos

que o Núcleo de Excelência que ora se propõe pode ser capaz de catalisar processo semelhante nas instituições parceiras. Para tal, a concepção do MASWeb contempla não apenas a intensificação das interações de pesquisa e desenvolvimento, buscando uma atuação mais próxima dos grupos das várias instituições, como investimentos para criar condições nas parceiras para a realização dos trabalhos de pesquisa de forma capilarizada.

Um outro ponto de destaque é a inserção e reconhecimento internacional do grupo de pesquisadores proponentes. Entre as várias evidências ressaltamos três. A primeira é o intenso intercâmbio de pesquisadores e alunos com instituições internacionais reconhecidas, as quais resultam em parcerias concretas em termos de publicações conjuntas e até co-orientações dos alunos. É ainda importante ressaltar que essas interações são diversificadas em termos de instituições e países (Estados Unidos, Canadá, Inglaterra, Espanha, França, Alemanha, Índia e Qatar, entre outros). A segunda evidência é a participação dos pesquisadores em comitês editoriais, comitês de programa e organização de eventos. Por exemplo, dois dos pesquisadores do grupo foram coordenadores gerais de eventos internacionais de relevância mundial. Nominalmente, Nivio Ziviani foi o coordenador da Annual International ACM SIGIR Conference 2005, realizada em Salvador em agosto de 2005, e Virgilio Almeida foi um dos coordenadores gerais da International World Wide Web Conference 2013, a maior e mais reputada conferência na área de Web, realizada no Rio de Janeiro em maio de 2013. A terceira evidência é não apenas o número de publicações geradas pelo grupo e a qualidade dos veículos e fóruns onde elas foram publicadas, mas o seu impacto, que soma mais de uma centena de milhar de citações no Google Scholar. Por exemplo, merecem destaque o livro Modern Information Retrieval (autoria de Ricardo Baeza-Yates e Berthier Ribeiro Neto), com mais de 12600 citações, e o livro Data Mining and Analysis: Fundamental Concepts and Algorithms (autoria de Mohammed Zaki e Wagner Meira Jr.) que, mesmo antes do seu lançamento em 2014 pela Cambridge University Press, registrou mais de 90000 downloads a partir do site dataminingbook.info.

O grupo é também responsável por diversas ações empreendedoras, como a criação das empresas Miner Technology Grouo, adquirida pelo Grupo UOL em 1999, e Akwan Information Technologies, adquirida pela Google Inc. em 2005, e do Centro de Tecnologia para a Web – CTWeb, recentemente instalado no BH-Tec. Em ambos os casos, foram propostos e implementados com sucesso modelos inovadores de participação da academia em empreendimentos de alto valor agregado incluindo recebimento de ações e compartilhamento de receitas.

Em termos de qualificação e experiência, assim como reconhecimento pelos pares, o grupo inclui em sua equipe 18 bolsistas de produtividade em pesquisa do CNPq, sendo 3 PQ1A, 1 PQ1C, 4 PQ1D e 10 PQ2. O grupo inclui ainda entre seus pesquisadores três membros titulares e três membros afiliados da Academia Brasileira de Ciências.

Desta forma, acreditamos que a institucionalização do MASWeb como grupo de excelência é o resultado natural de esforços exitosos de pesquisa, desenvolvimento tecnológico e formação nos últimos 30 anos, os quais podem alcançar patamares ainda mais significativos e relevantes com a formalização do Núcleo ora proposto.

2 Justificativa e Relevância

A Web é um fenômeno de mídia tão importante e revolucionário quanto foi a invenção da imprensa por Gutenberg no século XV. Enquanto a invenção da imprensa democratizou, em parte, o acesso à

informação, permitindo a disseminação em massa de livros e artigos impressos, a Web democratizou a *produção* de informação e conhecimento, até então monopólio de uns poucos agentes, tais como editoras e jornais. Com o advento da Web, qualquer pessoa com acesso a um computador e um mínimo de conhecimento pode produzir conteúdo que fica imediatamente acessível por meio de *hyperlinks* e pode ser posteriormente indexado e localizado através de máquinas de busca.

Esse fenômeno de democratização e popularização de produção de conteúdo ficou mais evidente com o surgimento de novas aplicações da Web, tais como YouTube, Facebook, e Flickr[1], que tornaram as tarefas de produção e publicação de conteúdo multimídia digital muito mais fácil, permitindo também maior interação (social) por meio desse conteúdo. De fato, estudos mostram que a Web está ocupando cada vez mais espaços tradicionalmente explorados por outros veículos tais como rádio e televisão [24]. Várias destas aplicações, tais como o Foursquare, o Twitter e o Google+[2], permitem o compartilhamento de conteúdo *georreferenciado*, fomentando novos tipos de interações entre as pessoas baseadas em aspectos relacionados à localização e à mobilidade. Por exemplo, o Foursquare permite que usuários compartilhem não somente sua localização em determinado momento por meio de *check-ins* em *venues*[3] registradas no sistema, mas também sua opinião sobre lugares visitados por meio das *tips*[4]. Para empresários e futuros clientes, as *tips* fornecem um *feedback* valioso, que pode determinar futuras ações [107]. Enfim, as novas aplicações da Web, além de servirem como veículo de comunicação, troca de conteúdo e trabalho colaborativo, abrem oportunidades para a criação de novas redes de discussão temáticas e de negociação, bem como novos serviços. Exemplos de serviços muito comuns na Web atualmente são serviços de recomendação, desde a recomendação de lugares e eventos até recomendação de amizades e colaborações acadêmicas [25, 78–80, 116].

Apesar de seu enorme sucesso, a Web apresenta inúmeros problemas, sendo que esse mesmo sucesso é responsável por uma grande parte deles. As pessoas[5], ao acessarem a Web, seja por meio das máquinas de busca, seja no contexto de novos serviços de caráter mais "social"e colaborativo, ainda encontram enormes dificuldades para alcançar seus objetivos informacionais [8]. Isso se deve não apenas à quantidade imensa de informação presente na Web, mas também à dificuldade de separar o material de interesse daquele de baixa qualidade ou irrelevante. Um outro aspecto que em certas situações exacerba o problema é o fato de a Web refletir, cada vez mais, o comportamento social das pessoas que dela fazem uso. Uma dessas reflexões negativas é constatada na presença de ações oportunistas e maliciosas, tais como Web *spamming,* poluição ativa de conteúdo, dentre outras.

A crescente popularização e participação das pessoas na geração de conteúdo e o seu caráter inerentemente social têm gerado também novos fenômenos e oportunidades que não têm sido devidamente exploradas por falta de um maior entendimento e compreensão desses fenômenos sociais. Por exemplo, o modelo econômico que sustenta boa parte da Web, baseado em propagandas *online*, tem tido sérias dificuldades para ser adaptado a sites baseados em redes sociais, e os motivos não são completamente conhecidos [72]. Além disto, a Web está em constante evolução. Logo, estudos sobre sua estrutura, aplicações e serviços devem levar em consideração este caráter inerentemente dinâmico para que soluções mais eficazes e robustas sejam desenvolvidas. Por exemplo, os fatores que impactam

[1]Respectivamente: http://www.youtube.com, http://www.facebook.com e http://www.flickr.com.

[2]Respectivamente: http://www.foursquare.com, http://www.twitter.com e http://plus.google.com.

[3]*Venues* correspondem a locais do mundo real, frequentemente locais comerciais, tais como uma loja ou um restaurante.

[4]Micro-revisões, limitadas a 200 caracteres, que usuários podem associar a *venues*.

[5]Os termos "pessoa" e "usuário" serão utilizados como sinônimos quando o contexto implicar "pessoas usando as informações e os serviços da Web".

a popularidade de conteúdo na Web é um tema de pesquisa bastante atual [44, 69, 87]. Um entendimento sobre tais fatores pode levar ao desenvolvimento de mecanismos de previsão de popularidade futura [83], que por sua vez podem ser explorados na otimização de vários serviços de recuperação de informação [50]. Por fim, mas não menos importante, o surgimento de novos serviços, dada a enorme taxa em que a Web vem crescendo, está limitado por aspectos de infraestrutura, tanto de software quanto de hardware, pois é necessário que essa infraestrutura atenda minimamente às demandas impostas pelas pessoas que as utilizam, em termos de desempenho, escalabilidade, segurança, confiabilidade e, mais amplamente, robustez, independentemente de tempo e localização geográfica. Portanto, o estudo dos diversos aspectos que envolvem a Web extrapola as atividades tradicionais de geração de conteúdo e de criação de novos serviços. O estudo da Web envolve o desenvolvimento de novas tecnologias que permeiam as suas diversas camadas e o entendimento da sociedade que utiliza os seus recursos [52].

Para lidar com os diversos problemas citados estamos propondo a criação do Núcleo de Excelência para a Web – Modelos, Algoritmos e Sistemas para a Web (MASWeb), cuja missão é aumentar a integração da Web com a sociedade. Para tal, partiremos de uma visão unificada da Web. Nessa visão, a Web é constituída por múltiplas camadas interdependentes de redes de relacionamentos complexas e dinâmicas, pelas quais a informação flui e é disseminada. Por rede, entendemos um conjunto de conexões ou relacionamentos entre pessoas, objetos informacionais (por exemplo, documentos), serviços e componentes de software e hardware. Essa visão permite uma abordagem unificada para atacar os desafios impostos pela Web nas diversas camadas propostas e o desenvolvimento de soluções sinérgicas, interdependentes e reusáveis através dessas camadas.

O projeto MASWeb está calcado em uma equipe multi-institucional de excelência e com competência comprovada para lidar com problemas relacionados à Web. Como será apresentado mais adiante, essa equipe reúne 33 pesquisadores de diferentes áreas da Computação, todos doutores, sendo 18 bolsistas de produtividade em pesquisa do CNPq. Além disso, os pesquisadores da equipe contam com numerosas colaborações e projetos em comum ao longo dos anos, bem como com parcerias com vários pesquisadores de instituições internacionais, também listados mais adiante neste documento. A origem de boa parte da integração entre os membros da equipe é o Instituto Nacional de Ciência e Tecnologia para a Web (InWeb), que reuniu pesquisadores da UFMG a parceiros de universidades em outros estados da federação e a pesquisadores de outras áreas. No âmbito do InWeb, foram formados diversos pesquisadores doutores, que agora atuam em universidades do estado de Minas Gerais, materializando o propósito original de disseminação da pesquisa em uma área de interesse estratégico para o país. Dessa forma, é natural o passo em direção à criação do Núcleo de Excelência aqui proposto, reunindo um grupo da UFMG, já enriquecido com novos professores contratados recentemente, a grupos da UFOP, UFJF, UFSJ, CEFET-MG e PUC Minas, constituídos em grande parte por egressos do Programa de Pós-Graduação em Ciência da Computação da UFMG. O MASWeb, portanto, avança no sentido de consolidar e formalizar efetivamente as colaborações já existentes que, pela intensidade e qualidade já verificadas, formam uma unidade coerente e capacitada para o alcance dos objetivos propostos. Além disso, a criação do Núcleo fortalecerá ainda mais a inserção internacional da equipe, oferecendo oportunidades especialmente importantes para os pesquisadores formados mais recentemente no sentido da interação com universidades, centros de pesquisa e institutos similares em outras partes do mundo. Nesse sentido, o MASWeb institucionalizará as fortes relações internacionais de seus membros participantes e permitirá a criação em Minas Gerais de um centro de referência para Web com presença internacional.

A geração de conhecimento e o domínio por cientistas e profissionais brasileiros de tecnologia avançada em áreas relacionadas à Web são de grande importância social e econômica, atingindo uma parcela importante da população de usuários de computadores e um grande número de empresas nacionais de diversos portes. O desenvolvimento de competência nacional nestas áreas é estratégico para o estado e para o país, pois abre perspectivas para que no futuro evitemos a importação ou até venhamos a exportar tal tecnologia. Minas Gerais, e Belo Horizonte em particular, tem tido destaque recente na criação de empresas de tecnologia da informação, que poderão se beneficiar tanto do avanço científico propriamente dito quanto da formação de recursos humanos altamente qualificados em nosso estado. Dadas as dimensões dos problemas envolvidos e a grande complexidade tecnológica para criação de novos negócios nesta área, o mercado para sistemas que permitam a utilização dos valiosos dados e informações que ocorrem em grande volume na Web é bastante promissor como fator de desenvolvimento econômico e social e de posicionamento estratégico do Brasil em um contexto global.

3 Programa de Pesquisa

Esta seção apresenta o programa de pesquisa do Núcleo de Excelência para a Web – MASWeb. Esse programa parte de uma visão unificada da Web. Essa visão permite atacar o objetivo de melhor integrar a Web com a sociedade por meio da investigação de três grandes desafios complementares, explorando as *expertises* e áreas de atuação dos membros da equipe. A visão unificada proposta é apresentada na Seção 3.1, enquanto os três grandes desafios derivados desta visão são discutidos na Seção 3.2. O caráter multidisciplinar dos desafios e programa de pesquisa propostos é discutido na Seção 3.3. Por fim, a Seção 3.4 apresenta os objetivos específicos que definem o programa de pesquisa do Núcleo, identificados como passos importantes na direção de se prover soluções efetivas para os três desafios apontados.

3.1 Visão Unificada da Web

A complexidade da Web vem de sua estrutura e do número muito grande de componentes (i.e., hosts, roteadores, servidores, serviços, hyperlinks e pessoas). Em uma visão micro, a Web é uma imensa infraestrutura de componentes, articulados através de linguagens e protocolos, fazendo uso de centenas de milhões de servidores e redes de comunicação [53]. Por outro lado, em uma escala macro, o uso coletivo dos componentes e serviços da Web cria outras redes superpostas à própria Web. Por redes, entendemos um conjunto de conexões ou relacionamentos entre pessoas, objetos informacionais, serviços e componentes de software e hardware. Portanto, neste projeto, a Web é vista como um sistema composto de múltiplas camadas de redes complexas dinâmicas e interdependentes, pelas quais a informação flui e é disseminada, conforme ilustrado na Figura 1.

O processo de disseminação de informação é iniciado na camada de interação (Figura 1(a)), onde as pessoas utilizam os serviços da Web para interagir e trocar conteúdo, produzindo e consumindo informação, e formando diversas redes de relacionamentos sociais. A camada de serviços (Figura 1(b)) é formada também por múltiplas redes uma vez que o conteúdo disponibilizado e materializado é potencialmente interconectado por meio de hyperlinks. Por exemplo, uma página no serviço YouTube pode conter links para serviços de blogs, sites de notícias (por exemplo, CNN) e outros serviços (por exemplo, Google). Além disso, cada serviço pode ser representado por uma rede de componentes de software incluindo servidores HTTP, servidores de aplicação e servidores de banco de dados. Por fim, a camada de infraestrutura (Figura 1(c)) é composta pela infraestrutura de rede de comunicação, incluindo

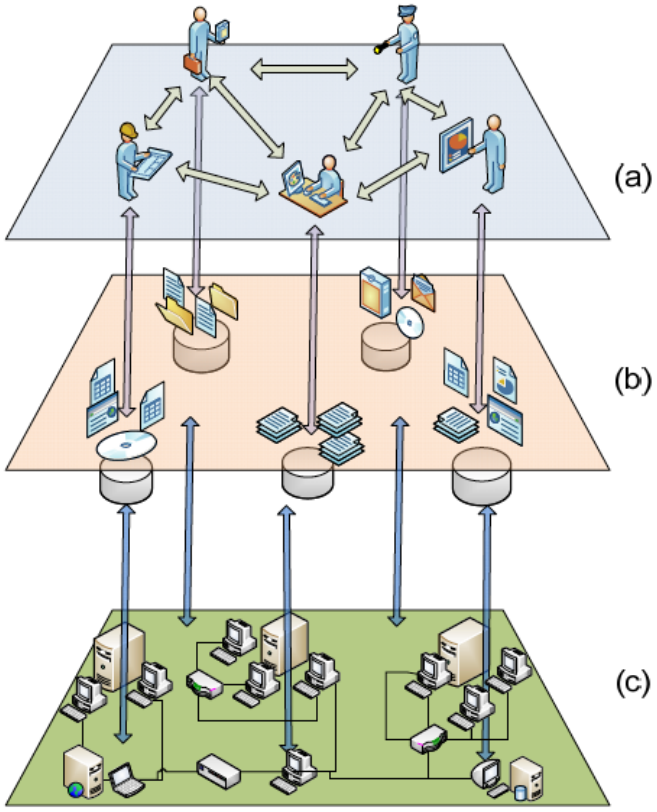

Figura 1: A Web como Múltiplas Camadas de Redes Complexas Dinâmicas e Interdependentes: (a) Camada de Interação, (b) Camada de Serviços e (c) Camada de Infra-estrutura.

não somente o hardware, mas também os componentes de software tais como protocolos para arquiteturas Peer-to-Peer (P2P) [10] e grades computacionais [67].

Todas essas redes são dinâmicas, evoluindo com o tempo, em resposta a como as pessoas se comportam, a falhas (propositais ou não) em componentes de software e hardware e à disponibilização de novos serviços e conteúdo. Elas são interdependentes, uma vez que alterações em qualquer camada podem causar impacto significativo nas demais. Por exemplo, a oferta de novos serviços tais como YouTube e Facebook (camada de serviços) levou ao surgimento de novos padrões de interação entre as pessoas, fomentando o estabelecimento de redes de relacionamentos baseadas em interesses (camada de interação). Em outro exemplo, a demanda cada vez maior das pessoas (camada de interação), distribuídas geograficamente, por serviços (camada de serviços) escaláveis, levou à popularização de arquiteturas descentralizadas (P2P e grades) como plataformas de implantação (camada de infraestrutura). Essas arquiteturas descentralizadas podem, por sua vez, afetar como as mesmas pessoas se comportam (camada de interação).

A partir dessa visão unificada da Web, o programa de pesquisa do MASWeb visa entender e explorar as várias redes de relacionamentos que constituem a Web, de forma a contribuir para melhorar a eficácia e a eficiência de seus serviços, principalmente serviços de disseminação de informação.

A melhoria da eficácia e da eficiência dos serviços da Web envolve tanto a melhoria da qualidade da informação que deve ser entregue de forma satisfatória, quanto a capacidade de prontamente disponibilizá-la para as pessoas que a solicitem, independentemente de tempo e localização geográfica.

Embora haja inúmeros esforços na literatura que visam propor soluções eficientes para atacar um ou outro aspecto das três camadas de redes de relacionamento [30, 35, 42, 42, 47, 49, 91], estes têm sido, em grande parte, isolados e independentes, o que leva a quatro conseqüências importantes: (1) a maioria dos resultados positivos obtidos tem sua contribuição limitada por enfatizar apenas um aspecto ou camada; (2) soluções são propostas partindo de premissas sobre o comportamento das pessoas e sobre o funcionamento de componentes que podem não refletir a realidade, levando a resultados que, na prática, podem ser bem inferiores; (3) muitas técnicas são aplicadas de forma separada e independente nas diferentes camadas da Web, quando uma solução mais abrangente e integrada poderia ser mais bem sucedida; e (4) soluções propostas separada e independentemente podem levar à melhoria de métricas diferentes, possivelmente conflitantes, limitando a eficiência da aplicação conjunta das mesmas. Assim sendo, fica clara a necessidade de uma abordagem mais unificada, como a proposta neste documento, aliando experiências complementares em uma rede de colaboração efetiva.

3.2 Desafios do Programa de Pesquisa

A partir da visão unificada elaborada acima, podemos vislumbrar três grandes desafios que devem ser atacados para o alcance do objetivo do núcleo de melhor integrar a Web à sociedade. O primeiro desafio é denominado "Identificação, Caracterização e Modelagem de Interesses e Padrões de Comportamento das Pessoas e das Redes Estabelecidas entre Elas na Web". Este desafio busca compreender e modelar aspectos das interações sociais que ocorrem entre as pessoas (camada de interação) por meio dos serviços (camada de serviços) e provê subsídios para entender tanto as necessidades das pessoas quanto o impacto que os padrões de comportamento e interações têm sobre as camadas inferiores. Uma vez compreendidas estas necessidades e padrões, o próximo grande desafio envolve o tratamento do conteúdo disponível na Web para lidar com seu enorme volume e heterogeneidade (por exemplo, em qualidade e formatos) de forma a melhor atender estas necessidades e acomodar os padrões de comportamento. Este desafio é denominado "Tratamento da Informação que Circula pelas Diversas Redes da Web". Por fim, a informação tratada deve ser entregue à pessoa que a solicitou de forma satisfatória, independente de tempo e lugar. Este último desafio, denominado "Entrega da Informação de Forma Satisfatória e Independente de Tempo e Lugar" trata do monitoramento e desenvolvimento de novas infra-estruturas para serviços Web.

Os três desafios descritos acima são elaborados a seguir.

3.3 Desafio 1: Identificação, Caracterização e Modelagem de Interesses e Padrões de Comportamento das Pessoas e das Redes Estabelecidas entre Elas na Web.

Este desafio envolve a identificação, caracterização e modelagem dos padrões de comportamento das pessoas ao interagirem por meio de serviços da Web bem como de aspectos das redes que emergem implícita e explicitamente a partir destas interações. O desafio busca determinar os interesses e necessidades das pessoas ao utilizarem os vários serviços da Web no dia-a-dia. Ele busca também identificar os padrões de comportamento e de interação entre estas pessoas que ocorrem mais frequentemente, bem como capturar o impacto que estes padrões possam ter nas decisões de projeto das camadas inferiores de serviços e de infra-estrutura. Em última instância, este desafio visa prover subsídios para o desenvolvimento de soluções que contribuam para a melhor eficiência e eficácia dos serviços da Web no atendimento das necessidades das pessoas.

O tratamento efetivo deste desafio envolve vários aspectos inovadores. Por exemplo, o comportamento das pessoas ao utilizar os vários serviços da Web requer novos modelos e técnicas de análise que capturem o comportamento coletivo que emerge da interação nos grupos e comunidades. Em particular, a interação entre pessoas mediada pela Web requer a adaptação de conceitos, tais como as regras de etiqueta online (*netiquette* [27, 86]). É importante também identificar, caracterizar e modelar como as características sociais dos usuários de serviços da Web, tais como reputação, confiança, egoísmo, cooperação e racionalidade [13] influenciam e são influenciadas pelo comportamento dinâmico das redes que compõem as camadas inferiores de serviços e de infraestrutura. De fato, vários dos problemas existentes em serviços da Web estão relacionados ao comportamento social (ou anti-social) das pessoas que os utilizam. Padrões de comportamento considerados oportunistas e maliciosos, tais como ataques de segurança, egoísmo e negação de serviço, disseminação de vírus e de conteúdo poluído, *spamming*, difamação e a auto-promoção são cada vez mais freqüentes, a despeito dos esforços para combatê-los [6, 21–23, 48, 73, 90, 109]. Estes padrões de comportamento e como eles são influenciados por aspectos das camadas de serviços e infra-estruturas podem ser mais bem investigados sob uma perspectiva multidisciplinar, à luz de elementos, modelos e teorias oriundos de diversas áreas como Psicologia, Comunicação e Ciências Sociais. Este é um claro exemplo onde a multidisciplinaridade pode levar a soluções inovadoras no avanço de problemas da Ciência da Computação.

Além disto, a importância da compreensão e caracterização do comportamento coletivo e das relações sociais entre as pessoas está também na identificação de agrupamentos (ou comunidades) temporais e geográficos, muitos dos quais compartilham interesses comuns e afinidades [85]. Alguns destes agrupamentos podem ser identificados a partir de redes implícitas de interesse e colaboração tais como a rede de co-autoria de publicações em uma comunidade científica, que podem ser extraídas de variadas fontes de informação disponíveis na Web. Agrupamentos de pessoas com interesses em comum e afinidades podem ser efetivamente explorados para melhorias de desempenho e eficácia de serviços da Web bem como no desenvolvimento de novos serviços (por exemplo, serviços de recomendação [25]).

A identificação, caracterização e posterior modelagem de padrões de comportamento e de redes de interação, incluindo aspectos como distribuições temporais e geográficas, padrões de visitação e de acesso, popularidade de objetos informacionais e de relacionamentos, passam ao longo dos seguintes eixos: (1) funcionalidade dos serviços (por exemplo: máquinas de busca, redes sociais, blogs), (2) tipos de objetos informacionais solicitados (por exemplo: texto, vídeo, áudio, código), (3) características dos usuários (por exemplo: idioma, referências geográficas) e (4) protocolos utilizados. Dada a escala da ordem de centenas de bilhões de componentes, novas técnicas experimentais serão necessárias para monitorar e coletar informações requeridas para caracterizações representativas e para o posterior desenvolvimento de modelos fidedignos.

Um aspecto importante a se considerar em qualquer estudo de caracterização e modelagem dos padrões de comportamento na Web é o caráter inerentemente dinâmico das aplicações neste ambiente, que implica que os padrões de comportamento e as redes de interação também evoluem com o tempo. Modelos que capturem aspectos chaves desse processo evolutivo são portanto necessários. Por exemplo, a popularidade dos conteúdos compartilhados em aplicações da Web, estimada pelo número de visualizações recebidas, reflete, em última instância, os interesses das pessoas. A investigação dos fatores que impactam como a popularidade de diferentes tipos de conteúdo evolui com o tempo pode levar ao desenvolvimento de novos modelos [69, 83]. Tais modelos, por sua vez, podem ser explorados para o desenvolvimento de serviços de recuperação de informação mais eficazes e robustos [50].

Mais ainda, a popularização de aplicações de compartilhamento de conteúdo geo-referenciado na Web, tais como o Foursquare, o Google+ e o Waze[6] abre oportunidades de se analisar padrões de mobilidade humana em uma escala nunca antes feita [97–100]. A partir da análise dos padrões de comportamento das pessoas que utilizam estas aplicações, pode-se inferir padrões de movimentação típicos em uma região alvo (por exemplo uma cidade) bem como padrões diferenciados decorrentes de eventos especiais. O potencial deste tipo de análise para o projeto de novos serviços (e.g., serviços de recomendação) bem como para suporte à tomada de decisões no planejamento urbano é enorme [14, 63].

A obtenção de dados reais na Web apresenta desafios por si só, como a preservação da privacidade e anonimidade das pessoas [54]. As limitações existentes para analisar e entender um sistema complexo como a Web abrem espaço então para o desenvolvimento de novos modelos, técnicas e algoritmos robustos a essas limitações. Além disto, a variedade de fontes e o enorme volume de objetos informacionais que podem servir como evidências sobre os padrões de comportamentos, de interações e, em última instância, sobre os interesses e necessidades das pessoas também apresentam desafios. Por exemplo, evidências podem ser extraídas a partir de registros (*logs*) mantidos pelos próprios serviços, de tráfego coletado em pontos selecionados da rede, de informações e páginas disponibilizadas nos serviços e de experimentos com usuários reais. Estas evidências precisam ser coletadas, armazenadas em repositórios e posteriormente processadas. Além disto, estes repositórios devem ser continuamente alimentados a partir do monitoramento freqüente ao longo do tempo.

Em suma, este desafio envolve a caracterização e modelagem do comportamento das pessoas, de suas interações, interesses e necessidades, a partir da coleta e processamento de uma multitude de evidências extraídas de diferentes fontes. Ele tanto se beneficia quanto fornece subsídios para o tratamento dos dois outros desafios (descritos abaixo). Como exemplo, o armazenamento, gerenciamento e processamento do enorme volume de dados referentes às evidências coletadas irá se beneficiar de várias técnicas e algoritmos desenvolvidos para atacar os desafios 2 e 3. Por outro lado, os novos modelos de comportamento do usuário que pretendemos produzir servirão de base para o desenvolvimento de modelos de carga e de tráfego mais realistas, que, por sua vez, alimentarão a avaliação experimental de mecanismos alternativos para os diferentes componentes das camadas de serviços e infra-estrutura (desafios 2 e 3). Além disto, a análise das diferentes redes que emergem das interações entre os usuários[7] podem auxiliar na determinação da influência de cada uma delas na disseminação efetiva de informação na Web. Por fim, o tratamento deste desafio irá não somente se beneficiar de conhecimento oriundo de outras Ciências, conforme discutido acima, mas também irá produzir conhecimento novo para entender o comportamento humano no mundo virtual da Web.

3.4 Desafio 2: Tratamento da Informação que Circula pelas Diversas Redes da Web

Uma vez identificados os interesses e padrões de comportamento dos usuários, a enorme gama de informação existente na Web precisa ser tratada de forma adequada, antes que essa informação possa ser entregue, diretamente ou através de serviços na Web, para atender aos interesses desses usuários. Por exemplo, é necessário remover conteúdo de baixa qualidade ou nocivo incluído por usuários maliciosos ou oportunistas, extraindo e organizando a informação de forma eficaz e eficiente. É também necessário

[6] http://www.waze.com

[7] Em um serviço de redes sociais como o YouTube, várias redes emergem implícita e explicitamente a partir das interações entre seus usuários. A rede de amigos, por exemplo, é estabelecida por relações explícitas de amizade entre os usuários. A rede de comentários, em contrapartida, emerge a partir da inserção de comentários de um usuário sobre um objeto (vídeo) compartilhado por outro.

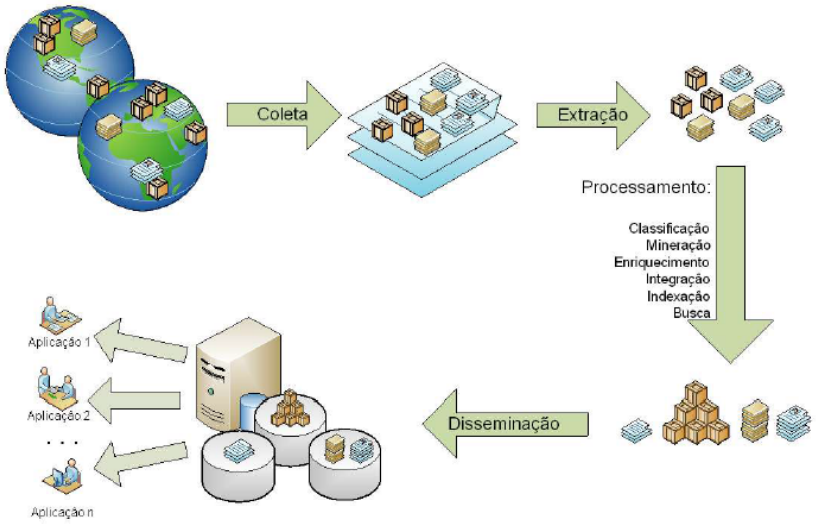

Figura 2: Ciclo de Coleta, Extração, Processamento e Disseminação de Informação Disponível na Web.

garantir que a informação obtida possa ser entregue de forma satisfatória no momento e no lugar em que o usuário a requisite.

Mais especificamente, por tratamento da informação entendemos o processo de coleta, extração, processamento e disseminação da informação existente na Web, como ilustrado na Figura 2. As dimensões atuais da Web, além da complexidade e da diversidade crescente de seus serviços, têm testado os limites das soluções atuais desenvolvidas para tratamento de informação em serviços convencionais, tais como técnicas de recuperação de informação em coleções controladas de documentos textuais e técnicas tradicionais de modelagem e armazenamento de bancos de dados.

Nesse sentido, novas soluções precisam ser desenvolvidas para tratar a enorme gama de informação de maneira escalável, eficaz e eficiente.

A tarefa de tratamento de informação na Web pode envolver, por exemplo, o desenvolvimento de coletores temáticos inteligentes [41], capazes de navegar por porções delimitadas das diversas redes da Web em busca de objetos informacionais (por exemplo, documentos) de um determinado domínio ou de áreas de interesse de uma determinada aplicação. Tais coletores seriam capazes de navegar dentro de determinados sites utilizando os conteúdos, estruturas e relacionamentos como guia para selecionar os objetos a coletar [15, 108]. Os objetos informacionais coletados passam, então, por um processo de extração, que consiste no processamento do seu conteúdo para a obtenção de dados que sejam de interesse da aplicação alvo [61]. O resultado desse processo de extração pode variar de acordo com a aplicação, podendo envolver a segmentação das páginas coletadas ou separação do seu conteúdo em categorias, ou simplesmente a geração de um repositório de dados em algum formato específico, como XML ou tabelas relacionais.

O resultado da extração pode ser ainda objeto de processamento adicional conforme os requisitos do serviço a ser construído. Por exemplo, o conteúdo extraído pode ser automaticamente classificado [28, 36] para a criação de diretórios temáticos, ser minerado para identificação de padrões e criação de novos repositórios [75], passar por um processo de integração [51] quando a extração é feita a partir de múltiplas fontes, ser enriquecido semanticamente por meio de anotações para determinação de seu significado ou, ainda, ser indexado para posterior utilização. Por fim, esse conteúdo deve ser entregue às pessoas por

meio das diversos serviços da Web. A entrega dessa informação de forma satisfatória, independente de tempo ou lugar é o terceiro grande desafio a ser atacado dentro do nosso programa de pesquisa.

3.5 Desafio 3: Entrega da Informação de Forma Satisfatória e Independente de Tempo e Lugar

Uma vez que se tenha conseguido produzir a informação desejada, o próximo grande desafio é a entrega dessa informação ao usuário de forma satisfatória. Este desafio envolve principalmente as duas camadas inferiores (Figura 1). Logo, o termo sistema será utilizado para se referir tanto a um serviço quanto à infra-estrutura de comunicação em que ele se apoia. Neste caso, este desafio inclui questões que vão desde a apresentação da informação através da interface do sistema até questões referentes ao impacto da camada de infra-estrutura sobre a comunicação entre usuários e serviços, envolvendo aspectos relativos a desempenho, escalabilidade e dependabilidade.

A interface de um sistema interativo é fundamental para o seu sucesso, uma vez que determina o uso que as pessoas podem fazer da funcionalidade oferecida. As novas possibilidades de acesso e comunicação oferecidas pela Web trazem desafios únicos para a interação dos usuários com os serviços. Na Web, o usuário tem acesso a uma quantidade cada vez maior de informação. Para que seja capaz de interpretar, selecionar e explorar grandes volumes de dados, formas distintas de visualização e exploração desses dados têm sido investigadas [31, 94]. Por exemplo, conhecendo-se de forma aproximada a localização geográfica do usuário [10], pode-se estimar o escopo geográfico de seu interesse, o que contribui para a seleção de conteúdo de relevância e para o estabelecimento de um contexto de uso. A possibilidade de comunicação e colaboração com outras pessoas requer que, além da usabilidade da interface, se atente também para a sua sociabilidade, ou seja, como a interface apoia os relacionamentos (curtos ou duradouros, geograficamente próximos ou a distância) entre pessoas e estimula ou desencoraja determinados tipos de comportamento dos usuários [84]. Além disso, o projeto de interfaces de sistemas deve levar em consideração outras questões de âmbito social tais como privacidade [105] e percepção sobre os outros usuários do sistema e tarefas por eles executadas [17].

No contexto da entrega da informação ao usuário, é também essencial que se avalie como a infra-estrutura de rede afeta a transferência de informação para o usuário e como os componentes de software podem melhor se organizar para realizar suas funções. A topologia da rede física sobre a qual a Web se apoia limita o volume de dados que pode ser trocado entre as partes de um sistema e cria pontos de interação muitas vezes inesperados entre usuários devido à competição por recursos da infraestrutura compartilhada. Sendo assim, faz parte desse desafio o desenvolvimento de soluções que facilitem o compartilhamento de recursos da rede, tais como técnicas para controle de contenção por recursos (p.ex., controle de congestionamento na rede [110]), arquiteturas distribuídas modernas que evitem a formação de tais pontos de contenção, como redes sobrepostas (overlays) [32], sistemas P2P [10] e grades computacionais (grids) [67]. Este desafio também inclui o desenvolvimento de técnicas que levem a uma melhoria da qualidade dos serviços oferecidos pela infra-estrutura para as camadas superiores. Soluções que enfatizem aspectos de dependabilidade, englobando noções de disponibilidade, confiabilidade, segurança, integridade e mantenibilidade em sistemas complexos [12, 88], são essenciais para a evolução da Web. Para isso pretendemos avaliar técnicas de medição, caracterização, análise, modelagem e projeto de sistemas que melhorem a dependabilidade dos sistemas envolvidos, bem como algoritmos e protocolos que permitam o desenvolvimento de soluções mais resistentes a falhas e a ataques.

Sob vários aspectos, ao focar a infra-estrutura da Web, este desafio cria paralelos importantes com os dois anteriores. Da mesma forma que o primeiro desafio aborda a caracterização e modelagem do comportamento de usuários, o presente desafio requer a caracterização e modelagem do comportamento da infra-estrutura, do tráfego entre aplicações, do impacto desse tráfego sobre a arquitetura da rede e vice-versa. Assim como o segundo desafio trata do processamento e armazenagem da informação extraída da Web, também aqui o problema de coleta e análise de grandes volumes de dados se faz presente, neste caso sob a ótica da comunicação entre os elementos da infra-estrutura de software e hardware [55]. Nesse sentido, esperamos que haja interação entre os três desafios na troca de experiências e na busca de soluções que possam ser aplicadas a ambos.

Finalmente, é importante lembrar que a infra-estrutura deve prover serviços computacionais e de comunicação para as demais camadas a partir de conjuntos de componentes complexos e heterogêneos. O panorama atual permite vislumbrar mudanças já em curso, em termos de arquitetura de sistemas, que permearão também a Web. Os sistemas multi-core, com dezenas ou mesmo milhares de núcleos de processamento por chip, muitas vezes heterogêneos, surgem trazendo a programação paralela como um elemento essencial a ser considerado. Com elas surgem novas possibilidades para melhoria de desempenho e controle do consumo de energia em grandes centros de processamento de dados e serviços [10]. Os recursos de virtualização permitem que projetistas desacoplem a noção lógica de elementos de processamento, como servidores e roteadores, dos dispositivos de hardware onde esse processamento será executado. Isso cria uma nova dimensão para a distribuição de serviços, balanceamento de carga e reconfiguração autonômica de sistemas [7,39,40,103]. Em ambos os contextos (sistemas multi-cores e sistemas virtualizados), é necessário o desenvolvimento de novos algoritmos, técnicas e métodos na área de sistemas que permitam oferecer aos usuários uma rede mais confiável, eficiente, flexível e fácil de usar.

3.6 Multidisciplinaridade

A presença maciça da Web na vida da sociedade moderna altera fundamentalmente as possibilidades individuais e coletivas, abrindo novas questões relacionadas a múltiplas áreas do conhecimento, como por exemplo, a difusão de inovação e conhecimento, os novos meios de comunicação (por exemplo, TV na Web, microblogs [8], redes sociais [9]), privacidade individual e segurança da sociedade e do Estado (por exemplo, crimes e ameaças eletrônicas). O entendimento do comportamento humano face à ampliação dos serviços eletrônicos é essencial para aproveitar os potenciais benefícios dos serviços da Web. O programa de pesquisa proposto para o Núcleo tem intrinsicamente um caráter multidisciplinar que deve ser explorado por pesquisadores de várias áreas. A Ciência da Computação e suas tecnologias têm uma horizontalidade que permeia praticamente todas as áreas da ciência. Em particular as áreas de Ciências Sociais e Humanas podem prover estrutura e fundamentação teórica para a compreensão da sociedade digital, baseada nas redes e na informação digital. Isso cria o contexto para os problemas centrais do MASWeb, nas áreas de classificação, armazenamento e recuperação de grandes massas de informação e conhecimento. Muitos dos problemas relevantes para a sociedade brasileira em áreas como saúde e educação são multidisciplinares nas possibilidades de solução. Um dos princípios do MASWeb é buscar direções inter e multidisciplinares, tendo em vista que problemas relevantes de pesquisa transcendem barreiras de disciplinas científicas tradicionais.

[8] *twitter -twitter.com*

[9] *facebook - www.facebook.com*, *instagram - instagram.com*

A complexidade deste tipo de pesquisa aumenta à medida que crescem o volume de dados e/ou os parâmetros a serem considerados. Por exemplo, a área de estudos globais do meio ambiente mostra a necessidade de integrar diferentes modelos: mudanças climáticas, sistemas humanos e naturais, desenvolvimento sócio-econômico e emissões de concentração de gases e poluentes [70]. O processo de expansão das redes, agregadas através da Web, leva a implicações ainda não estudadas nas várias dimensões da vida social. Por exemplo, em [26], há o seguinte cenário: "Por volta de 2047 ... toda informação sobre objetos físicos, incluindo humanos, edifícios, processos e organizações estarão online. Isso é ao mesmo tempo desejável e inevitável." O primeiro desafio do Núcleo de Excelência para a Web focaliza na direção da compreensão da interação entre grupos online e grupos sociais no mundo físico, que pode ser melhor investigada à luz de elementos e modelos oriundos das Ciências Humanas e Sociais.

Os outros desafios do MASWeb visam o tratamento das informações que circulam pelas diversas redes que compõem a Web e a respectiva entrega dessas informações às pessoas e usuários dos serviços. Isso requer competências multidisciplinares. Por exemplo, a estruturação e projeto de interfaces e serviços de acesso do cidadão a informação e conhecimento passam pela compreensão das diversidades e diferenças dos vários grupos sociais. O conhecimento das Ciências Humanas e Sociais é necessário, portanto, para os objetivos do MASWeb. Conceitos e teorias dessas ciências serão usados como elementos centrais na elaboração de algoritmos para recuperação, tratamento e entrega de informação, como por exemplo, reputação e egoísmo [89].

4 Objetivos e Metas

O objetivo do *Núcleo de Excelência para a Web* (MASWeb) é desenvolver modelos, algoritmos e novas tecnologias que permitam aumentar a integração da Web com a sociedade. Como resultado, nós esperamos tornar mais efetiva e mais segura a distribuição de informação, mais eficazes e eficientes os seus serviços, para que a Web se torne um vetor de mudanças sociais e econômicas no país.

As metas do MASWeb estão relacionadas com (i) a formação de recursos humanos qualificados, (ii) a produção de resultados de pesquisa inéditos que possam levar à criação de protótipos, os quais possam gerar tecnologias de ponta, e (iii) a disseminação de conhecimento para a sociedade. De forma mais específica temos como metas gerais:

1. Formar anualmente pelo menos 20 alunos de mestrado, 5 alunos de doutorado e orientar vários alunos em programas de Iniciação Científica;

2. Produzir impacto científico por meio da publicação anual de pelo menos 13 artigos em periódico e 40 artigos em congressos, todos de alta qualidade e visibilidade na comunidade científica internacional [10];

3. Realizar workshops com participação de todos os membros da equipe, seus alunos e outros membros da comunidade científica, para discussão e avaliação dos principais resultados obtidos bem como para definição de possíveis ajustes e novos direcionamentos;

[10] O maior número de artigos em conferência se deve ao fato de que, em Ciência da Computação, conferências assumem um papel de extrema importância para divulgação científica, dada a dinamicidade da área, que requer que os resultados de pesquisa sejam disseminados rapidamente. As conferências preenchem esse papel, sendo que as de primeira linha aceitam apenas trabalhos completos e têm uma taxa de aceitação ao redor de 15%, sendo mais competitivas que muitos periódicos de Computação. De fato, estudos mostram que nos principais departamentos de Computação do mundo a taxa é de aproximadamente 3 artigos em conferência para cada artigo em periódico publicado [60].

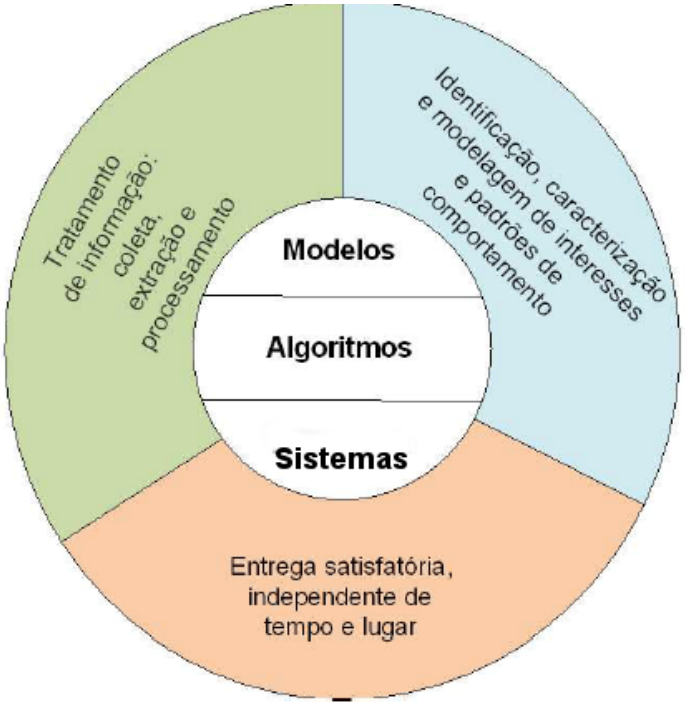

Figura 3: Objetivo Principal do Núcleo de Excelência para a Web: Modelos, Algoritmos e Tecnologia para a Web

4. Construir e manter atualizada uma biblioteca de componentes de software e protótipos para tratamento de objetos informacionais da Web (por exemplo, documentos, vídeos, imagens e áudios), com versões paralelas e distribuídas desses componentes. Essa biblioteca não somente viabilizará a execução da maioria das atividades de pesquisa a serem realizadas como também servirá de repositório para os resultados científicos e tecnológicos obtidos, que serão compartilhados com a comunidade científica, empresarial e a sociedade em geral;

Do ponto de vista científico, o núcleo MASWeb tem por objetivos e metas cobrir tópicos atualmente considerados grandes desafios de pesquisa em Ciência da Computação em vários países. Estes objetivos específicos são apresentados a seguir, agrupados em torno dos três principais desafios identificados na Seção 3.2.

4.1 Objetivos Específicos do Programa de Pesquisa

A Figura 3 ilustra a abordagem proposta, mostrando o objetivo principal do Núcleo de Excelência da Web de contribuir efetivamente para a maior integração da Web com a sociedade por meio de modelos, algoritmos e novas tecnologias que possam ser usados em soluções de problemas referentes aos três grandes desafios descritos acima. No que segue, definimos um conjunto de objetivos específicos a serem alcançados dentro do contexto de cada desafio. Esses objetivos específicos correspondem a etapas importantes a serem desenvolvidas para solucionar os desafios.

4.1.1 Objetivos Específicos do Desafio 1: Identificação, Caracterização e Modelagem de Interesses e Padrões de Comportamento das Pessoas e das Redes Estabelecidas entre Elas na Web

1. Caracterizar e modelar padrões de comportamento das pessoas e das redes estabelecidas a partir de suas interações para melhorar a eficiência e eficácia de serviços da Web.

2. Caracterizar e modelar padrões de interesse das pessoas a partir de redes implícitas de interesse e de colaboração identificadas e extraídas de fontes distintas e heterogêneas de informação presentes na Web.

3. Caracterizar e modelar a evolução temporal dos padrões de comportamento de usuários, particularmente em termos de como seus interesses e a estrutura das redes de interações evoluem com tempo, bem como explorar tais modelos no desenvolvimento de estratégias de previsão.

4. Identificar, caracterizar e modelar padrões de comportamento maliciosos, oportunistas e antisociais visando projetar mecanismos de detecção, controle e combate mais eficazes.

5. Modelar e desenvolver métodos que permitam a consideração e a apreciação de aspectos sociais gerados pelo impacto do uso dos serviços sobre as pessoas que os utilizam.

4.1.2 Objetivos Específicos do Desafio 2: Tratamento da Informação que Circula pelas Diversas Redes da Web

1. Desenvolver novas estruturas de dados e algoritmos para suportar a indexação e recuperação eficiente de itens de informação complexos, seu conteúdo, e seus relacionamentos com outros itens de informação.

2. Desenvolver coletores para fontes de dados heterogêneas (e.g., Web, bases de conhecimento, redes sociais online) a fim de construir coleções locais de dados complexos, representados com base nas estruturas de dados propostas.

3. Desenvolver mecanismos para a integração de dados provenientes de múltiplas fontes heterogêneas, de modo a unificar suas diversas e potencialmente conflitantes versões em itens de informação consolidados.

4. Desenvolver mecanismos para enriquecimento dos itens de informação consolidados, incluindo abordagens para reconhecimento de entidades, ligação semântica a bases de conhecimento, anotação, e classificação automática.

5. Desenvolver modelos e algoritmos de recomendação e recuperação de informação, mineração de dados e aprendizado de máquina para tarefas como inferir a relevância de itens de informação complexos dada uma necessidade de informação complexa, de modo a suportar tarefas analíticas, assim como tarefas de busca e recomendação.

4.1.3 Objetivos Específicos do Desafio 3: Entrega da Informação de Forma Satisfatória e Independente de Tempo e Lugar

1. Monitorar, caracterizar e modelar a infra-estrutura de redes sobre a qual se apoia a Web e as redes sociais, tanto no nível físico quanto de componentes de software (como topologia e capacidade de enlaces físicos ou plataformas de distribuição de conteúdo) a fim de se entender seu impacto sobre o desempenho das aplicações.

2. Explorar infra-estruturas de redes (overlays, caches, e centros de processamento na nuvem) para serviços baseados em redes sociais, máquinas de busca e distribuição de conteúdo.

3. Desenvolver ambientes para implementação, suporte em tempo de execução e análise de desempenho para algoritmos paralelos e distribuídos em arquiteturas multi-core, many-core e heterogêneas multi-nível.

4. Paralelizar algoritmos de recuperação de informação, mineração de dados e aprendizado de máquina para arquiteturas paralelas e distribuídas.

5. Investigar estratégias de interação que permitam aos usuários um melhor aproveitamento de serviços da Web no seu contexto, tais como personalização, programação pelo usuário final e visualização.

5 Metodologia de Desenvolvimento

Em projetos de longo prazo, como o proposto, envolvendo membros de múltiplas instituições, é necessário planejar com especial cuidado a metodologia aplicada na execução das atividades previstas para garantir o seu sucesso. Os membros da equipe já atuaram juntos em diversos projetos anteriores e apresentam histórico bastante positivo quanto à realização de projetos conjuntos, conforme já apresentado. Essas realizações anteriores são uma evidência de que a distância geográfica não é um obstáculo para a cooperação e a realização deste projeto. Além disso, as tecnologias atuais, tais como sistemas de e-mail e sistemas de videoconferência como o Skype, permitem reuniões com parceiros que estão fisicamente em diversos locais distintos.

A metodologia de trabalho proposta consiste de várias atividades administrativas, descritas a seguir. Estas atividades serão executadas em vários níveis, seguindo a estrutura organizacional e funcional do projeto. As atividades previstas são as seguintes:

- *Reuniões Técnicas: são reuniões de cunho essencialmente técnico, de periodicidade semanal ou quinzenal, com o objetivo de avaliar o desenvolvimento das tarefas e apresentar os progressos nas pesquisas realizadas. As reuniões técnicas deverão ocorrer de acordo com os temas de pesquisa, envolvendo pesquisadores e alunos participantes de cada tema. Essas reuniões podem envolver também membros da equipe com atuação em outros temas, sempre que o tema de pesquisa sendo explorado assim exigir. Essas reuniões tratarão de aspectos específicos dos problemas em estudo. Sempre que possível e se fizer necessário, membros da equipe que não se encontrarem fisicamente no mesmo local da reunião participarão remotamente, via videoconferência.*

- *Reuniões de Acompanhamento: são reuniões da coordenação do projeto com os pesquisadores responsáveis pelo desenvolvimento de cada tema, de periodicidade mensal, cujo objetivo principal*

será avaliar o desenvolvimento geral do projeto, corrigir eventuais distorções e discutir ações que visem manter as tarefas dentro do cronograma previsto. Sempre que necessário, sistemas de videoconferência poderão ser utilizados.

- *Reuniões de Avaliação: serão reuniões de periodicidade anual (ou menor, de acordo com a necessidade), com a participação presencial de todos os membros da equipe do projeto, pesquisadores e alunos, com o objeto de apresentar e discutir os resultados das pesquisas realizadas, verificar o desenvolvimento do projeto e realizar eventuais ajustes que se fizerem necessários.*

- *Workshops e Seminários: serão eventos abertos à comunidade para apresentação dos resultados parciais do projeto. Esses eventos também servirão como mecanismos de interação com empresas interessadas em formar parcerias para transformar resultados de pesquisa provenientes do projeto em tecnologias competitivas.*

O trabalho de pesquisa está organizado de forma hierárquica, com a designação de pesquisadores responsáveis para cada tema de pesquisa definido. Cada responsável por tema de pesquisa coordenará as atividades relativas ao tema, estabelecendo prazos, selecionando bolsistas e reportando à coordenação geral as atividades e resultados alcançados. A coordenação geral do projeto fará o acompanhamento do desenvolvimento geral em relação aos desafios propostos. A troca de informações entre pesquisadores ocorrerá de forma natural, pois o mesmo pesquisador poderá atuar em mais de um tema ou ainda coordenar uma tema e atuar como pesquisador em outros. Além disso, as reuniões de acompanhamento dos coordenadores de temas com a coordenação visam, além de avaliar o progresso da pesquisa, estabelecer as conexões entre as diversas iniciativas. As reuniões técnicas serão mais frequentes, por exemplo, semanais, para os membros de cada tema ou de temas fortemente relacionados. Note que as atividades do projeto são divididas por atividade de pesquisa e não por instituição, o que contribui para a integração entre os pesquisadores de instituições diferentes.

Estão também previstas visitas de intercâmbio envolvendo integrantes da equipe do projeto e de outros grupos de pesquisa internacionais que já cooperam com os pesquisadores das universidades participantes. O objetivo dessas visitas será fomentar a realização de trabalhos de pesquisa conjuntos e a troca de experiências com grupos internacionais que atuem em áreas correlatas às de competência da equipe do projeto. As publicações em conferências associadas a visitas técnicas serão incentivadas, com o objetivo de expandir ainda mais as relações internacionais do grupo.

6 Pesquisadores Participantes

A equipe principal deste projeto envolve pesquisadores de Ciência da Computação de seis instituições: Universidade Federal de Minas Gerais (UFMG), Universidade Federal de São João Del Rei (UFSJ), Universidade Federal de Ouro Preto (UFOP), Universidade Federal de Juiz de Fora (UFJF), Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e Pontifícia Universidade Católica de Minas Gerais (PUC-MG). A Tabela 1 lista os pesquisadores de cada instituição, com a descrição de cargo, caso pertinente, do nível da bolsa de produtividade em pesquisa do CNPq que possuem.

Em resumo, a equipe é composta por 33 pesquisadores, sendo todos doutores, 18 deles bolsistas de produtividade em pesquisa do CNPq. Dentre os bolsistas de produtividade, 3 são nível 1A, 1 é é nível

1C, 4 são nível 1D e 10 são nível 2. Além disso, desses pesquisadores 6 são membros da Academia Brasileira de Ciências (marcados com * na Tabela 1), sendo 3 titulares e 3 afiliados.

A equipe tem grande experiência na execução de projetos de pesquisa, tendo obtido sucesso com destaque para três aspectos importantes: (1) produção volumosa de artigos em conferências e periódicos internacionais de qualidade sendo que 6 pesquisadores possuem mais de 80 publicações indexadas pela DBLP[11]; (2) forte impacto na formação de recursos humanos com o envolvimento de grande número de alunos de graduação, mestrado e doutorado e (3) forte interação com o setor produtivo, levando à criação de novas empresas a partir de tecnologias desenvolvidas.

6.1 Colaboração entre Pesquisadores da Equipe

A motivação para a criação do *Núcleo de Excelência* para a Web (MASWeb) surgiu como uma maneira de formalizar o sucesso das colaborações de longa data entre as 6 instituições e os 33 pesquisadores da equipe. Essas colaborações podem ser avaliadas pelas publicações em conjunto em foruns de qualidade e visibilidade na comunidade científica, e podem ser melhor visualizadas por meio da rede de co-autoria dos pesquisadores ao longo dos últimos 5 anos (2009-2013) , mostrada na Figura 4. Nesta rede, uma aresta representa uma relação de co-autoria em pelo menos um artigo entre dois membros da equipe e as arestas são proporcionais ao total de publicações entre pares de membros da equipe. Como enfatizado nesta rede, os pesquisadores da equipe já vêm trabalhando em conjunto com sucesso, compartilhando e reutilizando conhecimentos. Portanto, a equipe já forma um grupo coeso, com expertises complementares comprovadas e com potencial para continuar a produzir avanços científicos e tecnológicos significativos, formalizado em um Núcleo de Excelência.

6.2 Parcerias com Pesquisadores Internacionais

A equipe mantém ainda parcerias acadêmicas com pesquisadores de diversas instituições internacionais que atuam em áreas relacionadas aos objetivos e metas do Núcleo. Citamos, em particular:

- Mohammed J. Zaki, Ressenlaer Polytechnique Inst.
- Srinivasam Parthasarathy, Ohio State University
- Joseph Konstan, University of Minnesotta
- Ricardo Baeza-Yates, Yahoo! Research
- Daniele Quercia, University of Cambridge
- Juliana Freire, NYU
- Gonzalo Navarro, Universidad de Chile
- Chedy Raissi, INRIA
- Krishna P. Gummadi, MPI
- Carlos Castillo, QCRI

[11] DBLP Computer Science Bibliography, http://www.informatik.uni-trier.de/ ley/db/. A DBLP é uma das mais completas e populares Bibliotecas Digitais de Ciência da Computação disponíveis na Web.

Coordenador	Nívio Ziviani
Pesquisadores	**UFMG**
Nivio Ziviani *	Prof. Titular, Bolsista de Produtividade em Pesquisa 1A
Virgilio A. F. Almeida*	Prof. Titular, Bolsista de Produtividade em Pesquisa 1A
Alberto H. F. Laender*	Prof. Titular, Bolsista de Produtividade em Pesquisa 1A
Wagner Meira Jr	Prof. Titular, Bolsista de Produtividade em Pesquisa 1C
Arnaldo A. Araújo	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 1D
Jussara M. Almeida*	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 1D
Marcos A. Gonçalves	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 1D
Renato A. C. Ferreira	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 1D
Adriano A. Veloso*	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Ana Paula Couto da Silva	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Clodoveu A. Davis	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Dorgival Guedes Neto	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Fabricio Benevenuto de Souza*	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Fernando M. Q. Pereira	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Gisele L. Pappa	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Mirella M. Moro	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Raquel C. Melo-Minardi	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Adriano C. M. Pereira	Prof. Adjunto
Berthier Ribeiro-Neto	Prof. Associado
Italo Fernando Scota Cunha	Prof. Adjunto
Loic P. G. Cerf	Prof. Adjunto
Olga N. Goussevskaia	Prof. Adjunto
Raquel O. Prates	Prof. Adjunto
Rodrygo Santos	Prof. Auxiliar
Pesquisadores	**UFSJ**
Leonardo Chaves Dutra da Rocha	Prof. Adjunto
Pesquisadores	**UFOP**
Anderson Almeida Ferreira	Prof. Adjunto
Luiz Henrique de Campos Merschmann	Prof. Adjunto
Pesquisadores	**UFJF**
Rodrigo Weber dos Santos	Prof. Adjunto, Bolsista de Produtividade em Pesquisa 2
Alex Borges Vieira	Prof. Adjunto
Pesquisadores	**CEFET-MG**
Evandrino G. Barros	Prof. Efetivo
Cristina Duarte Murta	Prof. Associado
Pesquisadores	**PUC-MG**
Humberto Torres Marques Neto	Prof. Adjunto
Wladmir Cardoso Brandão	Prof. Adjunto

Tabela 1: Pesquisadores Principais da Equipe.

- Mark Crovella, Boston University
- Mehdi Kaytoue, INSA de Lyon
- Amedeo Napoli, Loria-France
- Ponnurangam Kumaraguru, Indraprastha Inst. of Inf. Tech, India

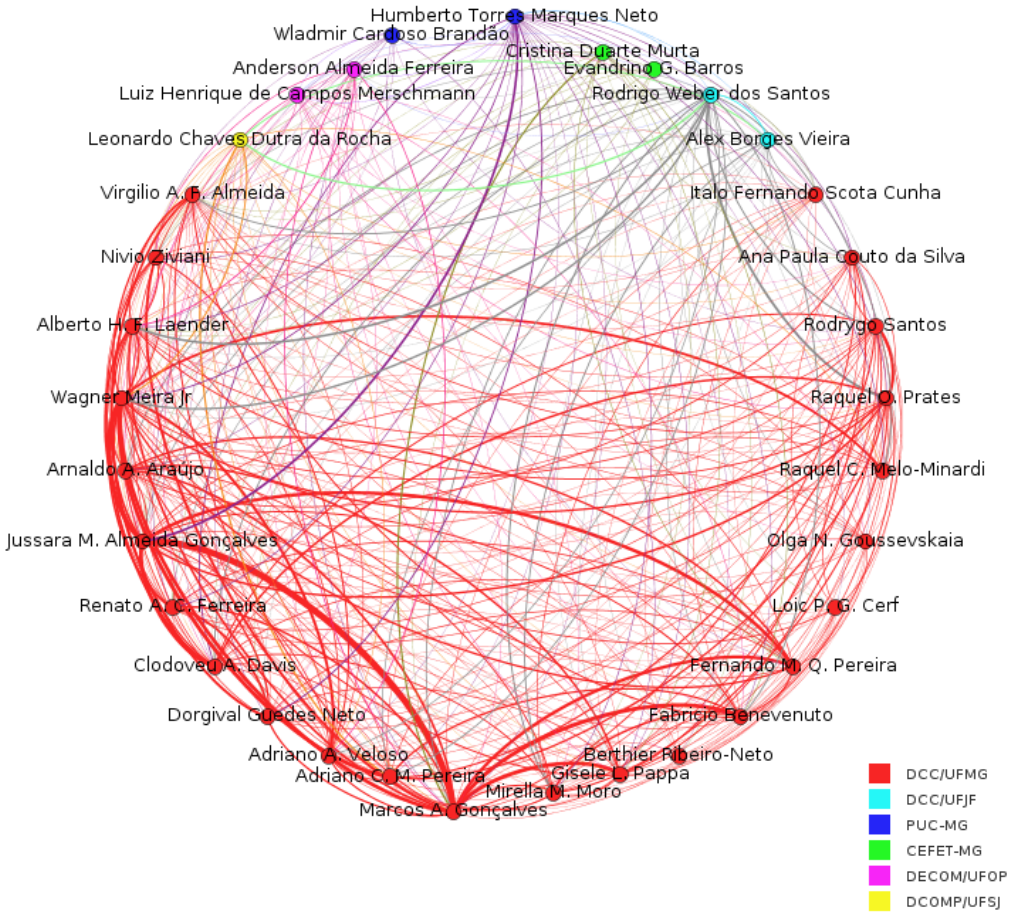

Figura 4: Rede de Co-Autoria da Equipe de Pesquisadores do MASWeb

- Sylvain Collange, Inria, FR
- Fabrice Rastello, ENS, Lyon, FR
- Mehdi Kaytoue, INSA, Lyon, FR
- Luciano Barbosa, IBM Brasil
- Frederico Fonseca, Penn State University, EUA
- Matei Ripeanu, University of British Columbia
- Ethan Katz-Bassett, University of Southern California
- Pável Calado, Instituto Superior Técnico - Universidade Técnica de Lisboa, Portugal
- Denilson Barbosa, University of Calgary, Canadá

- Mario Nascimento, University de Alberta, Canadá
- Azer Bestavros, Boston University, EUA
- Keith Ross, Polytechnic University, Brooklyn, EUA
- Vassilis J. Tsotras, University of California at Riverside, EUA
- Edward Fox, Virginia Tech, EUA
- Luis Bettencourt, Los Alamos National Institute, EUA
- Alex Freitas, University of Kent, Inglaterra
- Paulo Góes, University of Arizona – Eller School of Business, EUA
- Licia Capra, University College London, Inglaterra
- Marco Mellia, Politecnico di Torino, Italy
- Marco Ajmone, Politecnico di Torino, Italy
- Michela Meo, Politecnico di Torino, Italy

6.3 Publicações Selecionadas da Equipe

Esta seção relaciona algumas publicações dos pesquisadores da equipe em áreas relacionadas aos três grandes desafios abordados pelo MASWeb. Estas publicações atestam a qualificação e a complementaridade das pesquisas e a adequação da equipe ao programa de pesquisa proposto. Diversas publicações possuem como autores múltiplos pesquisadores da equipe, comprovando a sinergia já existente entre os pesquisadores participantes da proposta, ilustrada na Figura 4. Os pesquisadores da equipe estão marcados em negrito.

Publicações Relacionadas ao Desafio 1: Identificação, Caracterização e Modelagem de Interesses e Padrões de Comportamento das Pessoas e das Redes Estabelecidas entre Elas na Web.

1. Vasconcelos, Marisa ; **Almeida, Jussara M** ; **Goncalves, Marcos Andre** . What Makes your Opinion Popular? Predicting the Popularity of Micro-Reviews in Foursquare. In: ACM Symposium on Applied Computing, 2014, Gyeongju, Coréia do Sul. 29th ACM Symposium on Applied Computing, 2014.

2. Gonçalves, Glauber ; Drago, Idílio ; **Silva, Ana Paula Couto da** ; **Vieira, Alex Borges**; **Almeida, Jussara M** . Modeling the Dropbox Client Behavior. In: IEEE International Conference on Communications, 2014, Sydney, Australia. IEEE International Conference on Communications, 2014.

3. Santos-Neto, Elizeu ; Pontes, Tatiana ; **Almeida, Jussara M** ; Ripeanu, Matei . Towards Boosting Video Popularity via Tag Selection. In: Workshop on Social Multimedia and Storytelling, 2014, Glasgow, UK. Proc. Workshop on Social Multimedia and Storytelling, 2014.

4. Las-Casas, Pedro H.B. ; **Guedes, Dorgival**; **Almeida, Jussara M.** ; ZIVIANI, Artur ; **Marques-Neto, Humberto T.** . SpaDeS: Detecting spammers at the source network.Computer Networks (1999), v. 57, p. 526-539, 2013

5. Pinto, Henrique ; **Almeida, Jussara M.** ; **Gonçalves, Marcos André** . Using Early View Patterns to Predict the Popularity of YouTube Videos. In: 6th ACM International Conference on Web Search and Data Mining, 2013, Roma, Itália. Proc. 6th ACM International Conference on Web Search and Data Mining, 2013.

6. Martins, Eder F. ; Belem, Fabiano M. ; **Almeida, Jussara M.** ; **Gonçalves, Marcos** . Measuring and addressing the impact of cold start on associative tag recommenders. In: the 19th Brazilian symposium, 2013, Salvador. Proceedings of the 19th Brazilian symposium on Multimedia and the web - WebMedia '13. New York: ACM Press, 2013. p. 325-332.

7. Aggarwal, A. ; **Almeida, Jussara M** ; Kumaraguru, P. . Detection of Spam Tipping Behaviour on Foursquare. In: Mining Social Network Dynamics, in conjunction with World Wide Web Conference, 2013, Rio de Janeiro, Brazil. Proc. World Wide Web Conference, 2013.

8. Las-Casas, Pedro H.B. ; **Gonçalves, Marcos André** ; **Almeida, Jussara M.** ; **Marques Neto, Humberto T** ; **Guedes Neto, Dorgival** ; ZIVIANI, Artur . Adaptive Spammer Detection at the Source Network. In: IEEE Global Communications Conference, 2013, Atlanta, USA. Proc. IEEE Global Communications Conference, 2013.

9. Flores, A. G. ; **Vieira, Alex Borges** ; **Silva, Ana Paula Couto da**. ; Ziviani, Artur . Fast Centrality-Driven Diffusion in Dynamic Networks. In: SIMPLEX 2013, WWW 2013, 2013, Rio de Janeiro. 5th Annual Workshop on Simplifying Complex Networks for Practitioners - SIMPLEX 2013, WWW 2013, 2013.

10. Vasconcelos, Marisa; Ricci, Saulo ; **Almeida, Jussara M.**; **Benevenuto, Fabricio** ; **Almeida, Virgilio** . Tips, Dones and To-Dos: Uncovering User Profiles in FourSquare. In: ACM International Conference on Web Search and Data Mining (WSDM), 2012, Seattle, USA. Proc. ACM International Conference on Web Search and Data Mining (WSDM), 2012.

11. Gonçalves, Kênia Carolina ; **Borges, Alex** ; **Almeida, Jussara M.** ; **Silva, Ana Paula Couto**; **Marques Neto, Humberto T** ; Campos, Sérgio Vale Aguiar . Characterizing Dynamic Properties of the SopCast Overlay Network. In: 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, 2012, Garching, Alemanha. Proc. 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, 2012.

12. Pontes, T. ; Magno, G. ; Vasconcelos, Marisa ; Gupta, A. ; **Almeida, Jussara M.** ; Kumaraguru, P. ; **Almeida, Virgilio**. . Beware of What You Share: Inferring Home Location in Social Networks. In: International Workshop on Privacy in Social Data (PINSODA), 2012, Bruxelas, Bélgica. Proc. International Workshop on Privacy in Social Data, 2012.

13. Pontes, Tatiana ; Vasconcelos, Marisa ; **Almeida, Jussara M.** ; Kumaraguru, P. ; **Almeida, Virgílio** . We Know Where You Live: Privacy Characterization of Foursquare Behavior. In: Workshop on Location-Based Social Networks, in conjunction with 14th ACM International

Conference on Ubiquitous Computing, 2012, Pittsburg, EUA. Proc. Workshop on Location-Based Social Networks, 2012.

14. **Benevenuto, Fabrício** ; Rodrigues, Tiago ; **Veloso, Adriano** ; **Almeida, Jussara M.** ; Gonçalves, Marcos André ; Almeida, Virgilio A.F. . Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems. IEEE Transactions on Systems, Man and Cybernetics. Part B. Cybernetics, v. 42‘, p. 688-701, 2012.

15. Xavier, F. H. Z. ; Silveira, L. M. ; **Almeida, Jussara M.** ; Ziviani, Artur ; Malab, C. H. S. ; **Marques Neto, Humberto T** . Analyzing the Workload Dynamics of a Mobile Phone Network in Large Scale Events. In: First Workshop on Urban Networking, in conjunction with ACM CoNEXT 2012, 2012, Nice, França. Proceedings of the First Workshop on Urban Networking, 2012.

16. Figueiredo, Flavio ; **Benevenuto, Fabrício** ; **Almeida, Jussara M.** . The Tube over Time: Characterizing Popularity Growth of YouTube Videos. In: Proc. ACM International Conference on Web Search and Data Mining (WSDM), 2011, Hong Kong. Proc ACM International Conference on Web Search and Data MIning, 2011.

17. **Benevenuto, Fabrício** ; Rodrigues, Tiago ; **Almeida, Virgílio** ; **Almeida, Jussara** ; **Gonçalves, Marcos André** ; Ross, Keith . Video Pollution on the Web. First Monday (Online), v. 15, p. 1-13, 2010

18. **Gonçalves, Marcos André** ; **Almeida, Jussara M** ; Santos, L. G. P. ; **Laender, Alberto H.F.** ; **Almeida, Virgílio** . On Popularity in the Blogosphere. IEEE Internet Computing, v. 14, p. 30-37, 2010

Publicações Relacionadas ao Desafio 2: Tratamento da Informação que Circula pelas Diversas Redes da Web

1. Silva, R. ; **Gonçalves, Marcos André** ; **Veloso, Adriano**. A Two-stage active learning method for learning to rank. Journal of the Association for Information Science and Technology, v. 65, p. 109-128, 2014.

2. **Brandao, W. C.** ; Moura, E. S. ; **Santos, R. L. T.** ; Silva, A. S. ; **Ziviani, N.**. Learning to expand queries using entities. Journal of The American Society For Information Science and Technology (Online), 2014.

3. Botelho, Fabiano C. ; Pagh, Rasmus ; **Ziviani, Nivio**. Practical perfect hashing in nearly optimal space. Information Systems (Oxford), v. 38, p. 108-131, 2013.

4. De Carvalho, Moisés Gomes ; **Laender, Alberto H.F.** ; **Gonçalves, Marcos André** ; Da Silva, Altigran S.. An evolutionary approach to complex schema matching. Information Systems (Oxford), v. 38, p. 302-316, 2013.

5. Macdonald, C. ; **Santos, R. L. T.** ; Ounis, I. ; He, B.. About learning models with multiple query-dependent features. ACM Transactions on Information Systems, v. 31, p. 11:1-11:39, 2013.

6. Macdonald, Craig ; **Santos, Rodrygo L. T.** ; Ounis, Iadh. The whens and hows of learning to rank for web search. Information Retrieval (Dordrecht. Online), 2013.

7. Dalip, D. H. ; **Gonçalves, Marcos André** ; Cristo, Marco ; Calado, Pável. Exploiting User Feedback to Learn to Rank Answers in Q&A Forums: a Case Study with Stack Overflow. In: The Annual ACM SIGIR Conference on Information Retrieval, 2013, Dublin, Ireland. Proceedings of the 36th Annual ACM SIGIR Conference. New York: ACM Press, 2013.

8. Belem, F. ; **Santos, Rodrygo L.T.** ; **Gonçalves, Marcos André** ; **Almeida, J. M.**. Topic Diversity in Tag Recommendation. In: The ACM Conference Series on Recommender Systems (RecSys 2013), 2013, Hong Kong. Proceedings of the Seventh ACM Conference Series on Recommender Systems (RecSys 2013). New York: ACM, 2013.

9. Oliveira, D. M. ; **Laender, A. H. F.** ; **Veloso, A.** ; Silva, A. S.. FS-NER: A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. In: Third Workshop on Making Sense of Microposts, 2013, Rio de Janeiro. Proceedings of the 22nd International World Wide Web Conference (Companion Volume). New York: ACM, 2013. p. 597-604.

10. Lima, H. ; Silva, T. H. P. ; **Moro, M. M.** ; **Santos, R. L. T.** ; **Meira Jr., W.** ; **Laender, A. H. F.**. Aggregating Productivity Indices for Ranking Researchers across Multiple Areas. In: ACM/IEEE Joint Conference on Digital Libraries, 2013, Indianapolis, Indiana. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. New York: ACM, 2013. p. 97-106.

11. Lacerda, Anisio ; **Ziviani, Nivio**. Building user profiles to improve user experience in recommender systems. In: WSDM 2013, 2013, Rome, Italy. Sixth ACM International Conference on Web Search and Data Mining (Doctoral Consortium). New York: ACM, 2013. p. 759-764.

12. Menezes, D. ; Lacerda, A.M. ; Silva, L. ; **Veloso, A.** ; **Ziviani, N.**. Weighted Slope One Predictors Revisited. In: International World Wide Web Conference Companion, 2013, Rio de Janeiro. WWW 2013 Companion, 2013.

13. Lacerda, Anisio ; **Veloso, Adriano** ; **Ziviani, Nivio**. Exploratory and interactive daily deals recommendation. In: the 7th ACM conference, 2013, Hong Kong. Proceedings of the 7th ACM conference on Recommender systems - RecSys '13. New York: ACM Press. p. 439-16.

14. A. Bessa ; **Veloso, A.** ; **Ziviani, N.**. Using Mutual Influence to Improve Recommendations. In: 20th Symposium on String Processing and Information Retrieval, 2013, Jerusalem. 20th Symposium on String Processing and Information Retrieval. p. 17-28.

15. Santos, A. S. R. ; **Ziviani, N.** ; Almeida, J. ; Carvalho, C. ; Moura, E. S. ; Silva, Altigran Soares Da. Learning to Schedule Webpage Updates Using Genetic Programming. In: 20th Symposium on String Processing and Information Retrieval, 2013, Jerusalem. 20th Symposium on String Processing and Information Retrieval, 2013. p. 271-278.

16. **Rocha, L.** ; Salles, T. ; Mota, H. ; Mourao, F. ; textbfGonçalves, Marcos A. ; **Meira Jr, Wagner**. Temporal contexts: Effective text classification in evolving document collections. Information Systems (Oxford), v. 38, p. 388-409, 2013.

17. Carvalho, Moises Gomes De ; **Laender, Alberto H F** ; **Gonçalves, Marcos André** ; Silva, Altigran Soares Da. A Genetic Programming Approach to Record Deduplication. IEEE Transactions on Knowledge and Data Engineering (Print), v. 24, p. 399-412, 2012

18. Ribeiro, Marco Tulio ; Lacerda, Anisio ; **Veloso, Adriano** ; **Ziviani, Nivio**. Pareto-efficient hybridization for multi-objective recommender systems. In: the sixth ACM conference, 2012, Dublin. Proceedings of the sixth ACM conference on Recommender systems - RecSys '12. New York: ACM Press. p. 19-26.

19. Nascimento, C. ; **Laender, A. H. F.** ; Silva, A. S. ; **Gonçalves, M. A.**. A Source Independent Framework for Research Paper Recommendation. In: ACM/IEEE Joint Conference on Digital Libraries, 2011, Ottawa, Canada. Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. New York: ACM, 2011. p. 297-306.

20. Pereira, D. A. ; Ribeiro-Neto, B. A. ; **Ziviani, N.** ; **Laender, A. H. F.** ; **Gonçalves, M. A.**. A Generic Web-based Entity Resolution Framework. Journal of the American Society for Information Science and Technology (Print), v. 62, p. 919-932, 2011.

21. Assis, G. T. ; **Laender, A. H. F.** ; **Gonçalves, M. A.** ; Silva, A. S.. A Genre-Aware Approach to Focused Crawling. World Wide Web (Bussum), v. 12, p. 285-319, 2009.

Publicações Relacionadas ao Desafio 3: Entrega da Informação de Forma Satisfatória e Independente de Tempo e Lugar

1. B. R. Coutinho ; D. N. Sampaio ; **F. M. Q. Pereira** ; **W. Meira Jr**. Profiling divergences in GPU applications. Concurrency and Computation, v. 25, p. 775-789, 2013.

2. D. E. V. Pires ; L. C. Totti ; R. Moreira ; E. Fazzion ; O. Fonseca ; **W. Meira Jr** ; **R. C. Melo-Minardi** ; **D. O. Guedes Neto**. FPCluster: an efficient out-of-core clustering strategy without a similarity metric. Journal of Information and Data Management - JIDM, v. 3, p. 132-141, 2012.

3. **R. A. F. Ferreira** ; **D. O. Guedes Neto** ; **W. Meira Jr**. Anteater: Service-oriented data mining. In: Werner Dubitzky. (Org.). Data Mining Techniques in Grid Computing Environments. 1ed.Chichester, West Sussex: John Wiley & Sons, Ltd, 2008, v. , p. 179-199.

4. T. Ramos ; R. Oliveira ; A. P. Carvalho ; **R. A. F. Ferreira** ; **W. Meira Jr**. Watershed: A high performance distributed stream processing system. In: 23rd International Symposium on Computer Architecture, 2011, Vitória, ES. Proc. of the 23rd International Symposium on Computer Architecture, 2011. p. 191-198.

5. B. R. Coutinho ; D. N. Sampaio ; **F. M. Q. Pereira** ; **W. Meira Jr.** Performance debugging of GPGPU applications with the divergence map. In: 22nd International Symposium on Computer Architecture and High Performance Computing, 2010, Petropolis, RJ. Proc. of the 22nd International Symposium on Computer Architecture and High Performance Computing, 2010. p. 33-40.

6. **Í. Cunha** ; R. Teixeira ; D. Veitch ; and C. Diot. Predicting and Tracking Internet Path Changes. In: ACM SIGCOMM, Toronto, Canada, 2011.

7. G. L. M. Teodoro ; R. S. Oliveira ; O. Sertel ; M. Gurcan ; **W. Meira Jr.**; U. Catalyurek ; **R. A. F. Ferreira**. Coordinating the Use of GPU and CPU for Improving Performance of Compute

Intensive Applications. In: IEEE Cluster 09, 2009, New Orleans, EUA. Proc of IEEE Cluster 09, 2009.

8. Guimaraes, D. A. ; Arcanjo, F. L. ; Antuna, L. R. ; **Moro, Mirela M.**; **Ferreira, R. A. C.** Processing XPath Structural Constraints on GPU. Journal of Information and Data Management - JIDM, v. 4, p. 47-56, 2013.

9. Teodoro, George ; Hartley, Timothy D. R. ; **Ferreira, Renato** ; Catalyurek, Umit V. Optimizing dataflow applications on heterogeneous environments. Cluster Computing, v. 1, p. 1, 2011.

10. Andrade, G. ; Ramos, G. ; Madeira, D. ; Oliveira, Rafel Sachetto ; **Ferreira, R. A. C.** ; **Rocha, Leonardo**. G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering. In: International Conference on Computational Science, 2013, Barcelona.

11. Teodoro, George Luiz Medeiros ; Timothy D. R. Hartley ; Catalyurek, Umit ; **Ferreira, Renato** . Run-time Optimizations for Replicated Dataflows on Heterogeneous Environments. In: HPDC'2010, 2010, Chicago, IL. Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, 2010. p. 13-24.

12. Teodoro, George Luiz Medeiros ; Mariano, Nathan ; **Meira Jr., Wagner**; **Ferreira, Renato** . Tree Projection-Based Frequent Itemset Mining on Multicore CPUs and GPUs. In: SBAC-PAD 2010, 2010, Petrópolis, RJ. Proceedinds of the 22nd International Symposium on Computer Architecture and High Performance Computing, 2010. p. 47-54.

13. Silva, T. ; **Almeida, J. M.** ; **Guedes, D.** New strategies for Live P2P streaming based on the characterization of live user-generated video. Computer Networks, v. 5, p. 4055-4068, 2011.

14. Moraes, H. M. B. ; Nunes, R. V. ; **Guedes, D.** DCPortalsNg: efficient isolation of tenant networks in virtualized datacenters. In: ICN 2014, The Thirteenth International Conference on Networks, 2014, Nice, França,2014. p. 230-235.

15. Nunes, R. V. ; Pontes, R. L. ; **Guedes, D.** Virtualized Network Isolation Using Software Defined Networks. In: IEEE Conference on Local Computer Networks (LCN), 2013, Sydney. Proceedings of the 38th IEEE Conference on Local Computer Networks (LCN), 2013. p. 1-5.

16. Rodrigues, H. ; Santos, J. R. ; Turner, Y. ; Soares, P. V. ; **Guedes, D.** Gatekeeper: Supporting Bandwidth Guarantees for Multi-tenant Datacenter Networks. In: 3rd Workshop on I/O Virtualization, 2011, Portland, EUA, p. 1-7.

17. Lallo, P. S. ; **Pappa, G.** ; Augusto, H. ; **Guedes, D.** ; **Meira Jr., W.**. Map Ants, Reduce Work. In: 9th Metaheuristics International Conference (MIC 2011), 2011, Udine, Itália. Proceedings of the 9th Metaheuristics International Conference (MIC 2011), 2011.

18. Macambira, T. ; **Guedes, D.** A middleware for parallel processing of large graphs. In: 8th International Workshop on Middleware for Grids, Clouds and e-Science - MGC 2010, 2010, Bangalore, Índia. Proceedings of MGC 2010, 2010. p. 1-6.

19. G. Gonçalves ; A. Guimaraes ; **A. B. Vieira** ; **Í. Cunha** ; **J. M. Almeida**. Using Centrality Metrics to Predict Peer Cooperation in Live Streaming Applications. In: IFIP Networking, Prague, Czech Republic, 2012.

20. U. Javed ; **Í. Cunha** ; D. R. Choffnes ; E. Katz-Bassett ; A. Krishnamurthy ; T. Anderson. PoiRoot: Investigating the Root Cause of Interdomain Path Changes. In: ACM SIGCOMM, Hong Kong, China, 2013.

7 Cronograma das Atividades

Esta seção detalha as atividades previstas para execução deste projeto visando atingir cada um dos objetivos específicos listados na Seção 4.1. Ela também identifica os membros da equipe que estarão envolvidos em cada atividade. As atividades são agrupadas pelo desafio e pelo objetivo específico ao qual estão associadas. Um cronograma previsto para execução das atividades associadas a cada desafio é apresentado ao final da subseção correspondente.

7.1 Principais Atividades do Desafio 1

O desafio 1 consiste na identificação, caracterização e modelagem de interesses e padrões de comportamento das pessoas e das redes estabelecidas entre elas na Web. Cinco objetivos específicos foram identificados no contexto deste desafio, conforme apresentado na Seção 4.1. A seguir, as atividades previstas para se alcançar cada um destes objetivos são detalhadas.

Ressalta-se que a execução das atividades a seguir depende da disponibilidade de bases de dados que refletem as ações de usuários de diferentes aplicações. Para obter tais dados, pretendemos desenvolver ferramentas de coletas (e.g., ferramentas que utilizam a API das aplicações) e/ou utilizar ferramentas previamente desenvolvidas para coletar traços de uso e/ou de tráfego para as aplicações alvo do estudo bem como utilizar bases disponíveis publicamente (assim como feito anteriormente [19–21,44,83,107]).

7.1.1 Objetivo 1: Caracterizar e modelar padrões de comportamento das pessoas e das redes estabelecidas a partir de suas interações para melhorar a eficiência e eficácia de serviços da Web

As atividades previstas para se alcançar este objetivo têm duas frentes principais: padrões de comportamento de usuários de aplicações de compartilhamento na nuvem (particularmente Dropbox) e padrões de mobilidade de usuários inferidos de aplicações da Web geo-referenciadas. Pretendemos ainda explorar os padrões de comportamento identificados no desenvolvimento de novos serviços.

Atividade 1: Caracterização dos padrões de comportamento de usuários do Dropbox

Esta atividade consiste em identificar as principais variáveis que descrevem o comportamento dos usuários do Dropbox bem como caracterizá-las a partir de dados coletados de diferentes fontes. Em particular, pretendemos usar a ferramenta tstat [45] para coletar logs de tráfego de entrada e saída de vários pontos da Internet (e.g., universidades participantes deste projeto). Estes logs serão a principal fonte de dados para esta atividade. As variáveis identificadas devem capturar tanto aspectos do comportamento dos clientes (e.g., duração das sessões, tempos entre sincronizações sucessivas) quanto características dos conteúdos por eles compartilhados (e.g., tamanho e tipo dos arquivos).

Membros da equipe envolvidos nesta atividade: Jussara Almeida (UFMG), Ana Paula Couto da Silva (UFMG), Alex Borges Vieira (UFJF).

Atividade 2: Criação de um modelo de comportamento dos usuários do Dropbox

Esta atividade contempla a sumarização dos resultados de caracterização em um modelo de comportamento. Os diferentes padrões de comportamento identificados na atividade 1 deverão ser representados. Objetiva-se, em última instância, desenvolver um gerador de cargas sintéticas realistas e validá-lo usando logs de tráfego reais. O gerador poderá ser usado posteriormente para avaliar soluções existentes para armazenamento na nuvem (e.g., mecanismos implementados pelo Dropbox e por outras aplicações do mesmo tipo) bem como novas otimizações.

Membros da equipe envolvidos nesta atividade: Jussara Almeida (UFMG), Ana Paula Couto da Silva (UFMG), Alex Borges Vieira (UFJF).

Atividade 3: Caracterização de padrões de mobilidade de usuários inferidos de aplicações da Web

Nesta atividade pretendemos investigar o potencial de explorar dados coletados de diferentes aplicações da Web (por exemplo: Foursquare, Waze) para inferir padrões de mobilidade humana. Em particular, dando continuidade a uma pesquisa sendo desenvolvida em parceria com a Oi [112–114], nós pretendemos analisar o impacto de grandes eventos (e.g., Olimpíadas, Reveillon, um concerto) nos padrões de mobilidade humana. Até então, nossa pesquisa neste contexto tem focado na análise de dados de chamadas telefônicas (dados anonimizados). Pretendemos estender este estudo para incluir dados coletados de aplicações da Web.

Membros da equipe envolvidos nesta atividade: Jussara Almeida (UFMG), Clodoveu Davis (UFMG),Humberto Marques-Neto (PUC-Minas).

Atividade 4: Modelagem dos padrões de mobilidade

Esta atividade consiste no desenvolvimento de modelos que capturem os padrões de mobilidade humana. Pretendemos explorar modelos previamente propostos, como o SMOOTH [74], SLAW [64] e os modelos propostos em [5], e possíveis adaptações que se fizerem necessárias ou se mostrarem interessantes. Em particular, no contexto de grandes eventos, pretendemos desenvolver um modelo de simulação que represente a mobilidade humana em uma região alvo durante diferentes tipos de eventos. Este modelo poderá ser explorado, posteriormente, para suporte a tomadas de decisão e planejamento. Mais amplamente, os modelos de mobilidade que pretendemos desenvolver poderão ser aplicados, posteriormente, no desenvolvimento de sistemas de recomendação (e.g., recomendação de lugares [116]).

Membros da equipe envolvidos nesta atividade: Jussara Almeida (UFMG), Humberto Marques-Neto (PUC-Minas).

7.1.2 Objetivo 2: Caracterizar e modelar padrões de interesse das pessoas a partir de redes implícitas de interesse e de colaboração identificadas e extraídas de fontes distintas e heterogêneas de informação presentes na Web

Este objetivo está fortemente ligado ao objetivo 1, dado que os padrões de interesse e as redes estabelecidas entre os usuários são expressões de seu comportamento. Logo, assim como no objetivo 1, uma das frentes de trabalho abordará as aplicações de compartilhamento na nuvem, particularmente o Dropbox. Uma outra frente tratará das redes de colaboração científica. Por fim, uma terceira frente abordará a inferência de interesses de usuários para fins de personalização no contexto de recomendação de tags.

Atividade 1 - Modelagem da rede de compartilhamento de conteúdo no Dropbox

Nesta atividade iremos estudar os padrões de compartilhamento de conteúdo no Dropbox a partir de propriedades da rede implícita que emerge das interações de seus usuários. Métricas de redes complexas [82], tais como assortatividade e métricas de centralidade [76] serão analisadas. Além disto, a evolução da estrutura da rede ao longo do tempo será analisada. Um aspecto particular a ser investigado é a evolução das comunidades de usuários ao longo do tempo. Para tal utilizaremos algoritmos de detecção de comunidade estado-da-arte, tais como [58, 96].

Membros da equipe envolvidos nesta atividade: Ana Paula Couto da Silva (UFMG), Jussara Almeida (UFMG), Alex Borges Vieira (UFJF).

Atividade 2 – Caracterização das redes de colaboração científicas

Nesta atividade pretendemos analisar dados coletados de bibliotecas digitais acadêmicas, tais como LATTES e DBLP, visando caracterizar diferentes propriedades da rede de colaboração entre autores com foco particular na influência dos pesquisadores. Alguns aspectos que pretendemos analisar são: quais são os pesquisadores mais influentes e a intensidade de colaboração entre diferentes pesquisadores. A partir da caracterização inicial, novas métricas de caracterização de pesquisadores influentes poderão ser propostas.

Membros da equipe envolvidos nesta atividade: Ana Paula Couto da Silva (UFMG).

Atividade 3 – Modelagem do processo de formação de parcerias em redes de colaboração científicas

Através da definição de redes de coautoria, nesta atividade iremos estudar como grupos de pesquisa são formados e se existe preferências de estabelecimento destas parcerias, baseando-se na informação de artigos publicados conjuntamente. Para tal, iremos aplicar o algoritmo proposto em [96].

Membros da equipe envolvidos nesta atividade: Ana Paula Couto da Silva (UFMG).

Atividade 4 - Desenvolvimento de métricas que capturam os interesses dos usuários para fins de personalização da recomendação de *tags*.

Esta atividade vem complementar resultados recentes de membros da equipe no desenvolvimento de mecanismos de recomendação de *tags*. Os mecanismos desenvolvidos até então [18–20, 68]

recomendam *tags* relevantes, novas e diversificadas para um conteúdo alvo (e.g., um vídeo) sem levar em consideração o perfil do usuário alvo. Buscamos agora desenvolver métricas que capturam os interesses e perfis dos usuários para posteriormente explorá-las no desenvolvimento de métodos de recomendação personalizada aos interesses do usuário alvo. Uma abordagem possível consiste em capturar os interesses dos usuários por determinados tópicos em função das *tags* usadas por eles. Neste caso, a métrica frequência de uso de diferentes *tags*. Outras abordagens que deverão ser consideradas podem explorar as redes de relacionamentos entre diferentes usuários.

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG), Marcos André Gonçalves (UFMG).

Atividade 5- Validação das métricas de interesses em mecanismos de recomendação de *tags*.

Nesta atividade, pretendemos validar as métricas de interesses propostas estendendo os métodos de recomendação de *tags* previamente propostos para incluí-las. O impacto da personalização na qualidade das *tags* recomendadas será avaliado em bases de dados coletados de diferentes aplicações (disponíveis em nossos laboratórios [68]).

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG), Marcos André Gonçalves (UFMG).

7.1.3 Objetivo 3: Caracterizar e modelar a evolução temporal dos padrões de comportamento de usuários, particularmente em termos de como seus interesses e a estrutura das redes de interações evoluem com tempo, bem como explorar tais modelos no desenvolvimento de estratégias de previsão

O foco deste objetivo será em caracterizar e modelar a popularidade do conteúdo compartilhado em aplicações de redes sociais, principalmente YouTube e Foursquare. Nestas aplicações, a popularidade de um conteúdo (e.g., um vídeo) pode ser afetada por uma multitude de fatores incluindo características do usuário que criou o conteúdo e de sua rede de interação, bem como fatores externos à aplicação. Mais ainda, a popularidade de um conteúdo pode ser estimada por diferentes medidas, dependendo do interesse particular. De fato, a grande maioria dos estudos sobre popularidade de conteúdo online abordam a popularidade estimada pelo número total de visualizações [4, 29, 37, 69, 77, 83, 87, 104]. Entretanto, o número de usuários únicos (audiência) que visualizaram o conteúdo pode ser uma medida mais interessante se a popularidade for explorada para serviços de propaganda. Logo, o estudo de popularidade de conteúdo sob diferentes perspectivas (i.e., diferentes medidas) é uma parte inerente a todas as atividades associadas a este objetivo, listadas a seguir. Vale ressaltar que estas atividades serão realizados utilizando várias bases de dados que já temos disponíveis em nossos laboratórios, previamente coletadas dos YouTube e do Foursquare.

Atividade 1: Caracterização da importância relativa de vários atributos para a popularidade de um conteúdo.

Nesta etapa iremos identificar e selecionar um número de fatores (ou atributos) que podem ser relevantes para a popularidade de um conteúdo e quantificar a sua importância relativa utilizando

métodos estatísticos, tais como correlação e modelos de regressão. Esta caracterização já foi realizada previamente por nós considerando como medida de popularidade o número total de visualizações [44]. Pretendemos dar continuidade à pesquisa já realizada focando agora em outras medidas de popularidade, particularmente o número de usuários distintos, ou audiência, do conteúdo.

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG).

Atividade 2: Caracterização dos padrões de evolução de popularidade

Pretendemos caracterizar como a popularidade de um conteúdo evolui com o tempo respondendo questões como: *as curvas de popularidade exibem um pico claro? Se sim, em quanto tempo, desde sua criação, o conteúdo atinge o seu pico de popularidade?* Pretendemos identificar padrões de evolução de popularidade usando algoritmos de agrupamento de séries temporais, tais como o algoritmo K-Spectral Clustering recentemente proposto [115]. Pretendemos ainda correlacionar os padrões identificados com os atributos previamente selecionados para caracterizar os tipos de conteúdos que exibem cada padrão. A comparação dos resultados obtidos para diferentes medidas de popularidade é de particular interesse.

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG).

Atividade 3: Desenvolvimento de modelos de evolução de popularidade de um conteúdo *online*

Pretendemos avaliar a eficácia de diferentes modelos disponíveis na literatura para *descrever* a evolução da popularidade de um conteúdo [37, 69, 115]. Muitos destes modelos assume a existência de um único pico relevante na curva de popularidade [37, 69]. Pretendemos avaliar até que ponto isto é válido em nossas bases de dados. Pretendemos ainda propor novos modelos que sejam mais precisos e adequados para cada tipo de conteúdo, levando em consideração diferentes medidas de popularidade. Em particular, o papel da revisita (i.e., múltiplas visualizações de um mesmo conteúdo por um mesmo usuário) na evolução de popularidade de um conteúdo, até então não abordado em trabalhos anteriores, deverá ser investigado.

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG).

Atividade 4: Desenvolvimento de modelos de previsão de popularidade.

Nesta etapa, pretendemos desenvolver modelos de previsão de popularidade, estendendo resultados recentes [83]. No contexto específico do YouTube, pretendemos desenvolver modelos para prever o padrão de evolução da popularidade de um conteúdo. Esta frente de pesquisa é motivada por resultados recentes que indicam que tais previsões podem melhorar significativamente as previsões de popularidade para uma data futura [83]. Especificamente, pretendemos desenvolver estratégias para prever, após um certo período de monitoração, qual padrão de evolução a popularidade de um vídeo irá seguir, assumindo um conjunto de padrões previamente identificados. A identificação destes padrões será feita em etapa anterior (Atividade 2), usando, por exemplo, algoritmos de agrupamento de séries temporais [115]. Assim, a tarefa de prever o padrão de evolução é mapeada na tarefa de classificar os vídeos de entrada em diferentes classes, cada uma representando um padrão previamente observado.

Os modelos de previsão de padrão de evolução serão posteriormente explorados no desenvolvimento de novos modelos de previsão de popularidade, a partir da criação de modelos especializados para

cada padrão identificado [83]. Os modelos desenvolvidos serão comparados com modelos alternativos disponíveis na literatura (e.g., [4, 83, 87].

Membros da equipe envolvidos nesta atividade: Jussara Marques de Almeida (UFMG) e Marcos André Gonçalves.

7.1.4 Objetivo 4: Identificar, caracterizar e modelar padrões de comportamento maliciosos, oportunistas e antissociais visando projetar mecanismos de detecção, controle e combate mais eficazes

Cada vez mais dados de redes sociais são utilizados para a construção de novas aplicações. Entretanto, serviços baseados nesse tipo de dados estão vulneráveis a diferentes formas de ataques e manipulações [48, 71].

O foco deste objetivo será identificar, caracterizar e modelar padrões de comportamento maliciosos, oportunistas e sociais, principalmente em sistemas como o Twitter, YouTube e FourSquare, visando projetar mecanismos de detecção, controle e combate a esses comportamentos. Em sistemas como o Twitter, contas faltas e robôs infiltrados podem não apenas divulgar Spam ou conteúdo indesejado, mas também visar a manipulação de estatísticas e ferramentas que exploram dados de redes sociais, como monitores sobre a repercussão de eleições ou qualquer outro evento. Por outro lado, sistemas como o FourSquare e o Yelp precisam lidar com inúmeras tentativas de enviesar a opinião de outros usuários através da postagem de revisões positivas ou negativas falsas sobre lugares. Cabe ressaltar que estas atividades serão realizadas utilizando várias bases de dados que já temos disponíveis em nossos laboratórios, previamente coletadas em trabalhos anteriores.

Atividade 1: Investigar o impacto da infiltração de robôs em redes sociais.

Nesta atividade visamos investigar como ocorre a infiltração de Robôs e contas faltas em redes sociais, em particular no Twitter, buscando identificar padrões de conexão e comunicação dessas contas a usuários reais das redes sociais ou mesmo a identificação de ataques coordenados, com a criação de múltiplas contas spam.

Membros da equipe envolvidos nesta atividade: Fabrício Benevenuto (UFMG).

Atividade 2: Investigar a manipulação da opinião dos usuários de sistemas sociais na Web.

Nesta atividade pretendemos investigar se é possível manipular opinião dos usuários de sistemas sociais na Web através da postagem automática de revisões e comentários associados a objetos na Web, como por exemplo, vídeos no YouTube, locais no FourSquare e no Yelp, livros na Amazon, etc. Com isso pretendemos investigar se ataques que visam manipular revisões e comentários associados a objetos na Web conseguem realmente ser bem sucedidos em manipular a opinião de usuários sobre o objeto. Pretendemos ainda explorar as formas de ataques realizadas em revisões/comentários associadas a objetos de mídias, visando identificar padrões de comportamento associados à postagem desses comentários capazes de diferenciar postagens que visam manipular opiniões de outros usuários.

Membros da equipe envolvidos nesta atividade: Fabrício Benevenuto (UFMG) e Jussara M. Almeida (UFMG).

Atividade 3: Detecção e combate a usuários maliciosos.

Nesta atividade pretendemos investigar estratégias de detecção e combate a usuários que realizam ataques identificados nas etapas anteriores, buscando soluções calcadas em diferentes técnicas, como o uso do grafo de conexões entre os usuários [48] ou mesmo aprendizagem de máquina [21].

Membros da equipe envolvidos nesta atividade: Fabrício Benevenuto (UFMG) e Adriano Veloso (UFMG).

7.1.5 Objetivo 5: Modelar e desenvolver modelos e métodos que permitam a consideração e a apreciação da qualidade de uso do sistema e também de aspectos sociais gerados pelo impacto do uso dos serviços sobre as pessoas que os utilizam

Sistemas colaborativos na Web 2.0, sejam voltados para o trabalho ou para relações sociais, trazem à tona diversos aspectos que podem surgir em consequência das relações que promovem entre os usuários, como por exemplo privacidade ou reputação [3, 118]. O foco deste objetivo é na proposta e avaliação de modelos e métodos que permitam a análise de aspectos sociais no sistema, sob o ponto de vista tanto do projetista do sistema, quanto dos seus usuários. Sob a perspectiva do projetista o foco é permitir que ele antecipe como as definições sobre a interação no sistema podem promover ou não determinados aspectos sociais de interesse ou mesmo avaliar o sistema (durante o design) em relação a eles. Em relação aos usuários, o foco é na análise do impacto social do sistema sobre eles [66].

Atividade 1: Geração de modelo de análise de projeto de configurações.

A grande variedade de usuários e formas de utilizar sistemas colaborativos na Web 2.0 normalmente é contornando oferecendo aos usuários a flexibilidade de customizar alguns aspectos do sistema. Pretendemos gerar um modelo de design que permita ao projetista de sistemas colaborativos descrever as configurações que pretende oferecer aos usuários no sistema e antecipar os cenários de uso que poderão ser criados pelos usuários e seus impactos.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

Atividade 2: Geração de um modelo descritivo de elementos de sociabilidade.

Pretendemos através de uma pesquisa analítica identificar os elementos que são relevantes para influenciar aspectos da sociabilidade [46] a ser gerada a partir do uso de um sistema colaborativo. As dimensões de sociabilidade identificadas poderão ser utilizadas tanto para caracterizar qualitativamente diferentes sistemas, como poderão apoiar o desenvolvimento e avaliação desses sistemas.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

Atividade 3: Análise de métodos de avaliação de qualidades de uso em sistemas colaborativos.

Atualmente existem diversos métodos propostos para avaliação de sistemas colaborativos. Muitas vezes eles têm focos distintos e adotam técnicas também diferentes [11]. Poucos destes métodos estão bem consolidados. Nosso objetivo é fazer uma análise dos diferentes métodos existentes, e identificar e

avaliar aqueles que se propõem a analisar como os aspectos sociais são apresentados e na interface do sistema e seus impactos sociais na interação.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

Atividade 4: Análise interdisciplinar de privacidade em redes sociais.

Redes sociais apresentam diversos desafios em relação ao controle de privacidade que os seus usuários podem ter sobre suas próprias informações. O uso disseminado dos sistemas de rede social na sociedade atual impacta a forma das pessoas perceberem e tratarem a troca de informação com outras pessoas. Assim, nosso objetivo é fazer uma análise ampla que parta da análise de como os sistemas de redes sociais permitem e apresentam questões de privacidade aos seus usuários, até a análise de como isso impacta as relações entre as pessoas. Para isso será feito um estudo em parceria com pesquisadores da área de Sociologia e Estatística.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

7.1.6 Cronograma de Execução das Atividades do Desafio 1

A seguir é apresentado um cronograma previsto para execução das atividades descritas nas seções 7.1.1–7.1.5.

7.2 Principais Atividades do Desafio 2

O desafio 2 contempla a representação, descoberta, reconciliação, enriquecimento, e recuperação da informação presente em redes complexas na Web, em consonância com os cinco objetivos específicos descritos na Seção 4. No restante desta seção, detalhamos as atividades de pesquisa previstas para a realização de cada um desses objetivos.

7.2.1 Objetivo 1: Desenvolver novas estruturas de dados e algoritmos para suportar a indexação e recuperação eficiente de itens de informação complexos, seu conteúdo, e seus relacionamentos com outros itens de informação.

Este objetivo visa ao projeto e implementação de um sistema base para recuperação de informação em redes complexas na Web. Esse sistema deve suportar a representação de itens de informação com conteúdo diverso, bem como a possibilidade de relacionamentos de tipos diversos entre esses itens. O sistema deve atender a requisitos de aplicações interativas, como busca e recomendação, bem como de aplicações analíticas, como mineração de dados.

Atividade 1: Modelagem da estrutura base para representação de informação.

Esta atividade contempla a modelagem conceitual de uma estrutura de dados expressiva para o armazenamento de itens de informação, independentemente do esquema de dados subjacente a esses itens. Essa estrutura deverá possibilitar o acesso eficiente a esses itens, seja por meio de seus identificadores, seja por meio de outras características desses itens. Finalmente, essa estrutura deverá possibilitar também o acesso eficiente a itens relacionados a um dado item, por meio de múltiplas relações heterogêneas.

Membros da equipe envolvidos nesta atividade: Alberto Laender (UFMG), Rodrygo Santos (UFMG), Anderson Ferreira (UFOP), Luiz Henrique Merschmann (UFOP), Humberto Marques Neto (PUCMG), Wladmir Brandão (PUCMG).

Atividade 2: Desenvolvimento de mecanismos de indexação e recuperação.

Esta atividade visa ao desenvolvimento de mecanismos eficientes para indexação e recuperação de dados a partir da estrutura de dados proposta na Atividade 1. Tais mecanismos deverão possibilitar a indexação e a recuperação de novos dados em tempo real. Além disso, para suportar o processamento de dados massivos, esta atividade contempla também a investigação de modelos de particionamento e replicação de dados, além de políticas de caching em múltiplos níveis, com vistas a eficiência de tempo e espaço, escalabilidade, e disponibilidade.

Membros da equipe envolvidos nesta atividade: Nivio Ziviani (UFMG), Rodrygo Santos (UFMG), Dorgival Guedes (UFMG).

7.2.2 Objetivo 2: Desenvolver coletores para fontes de dados heterogêneas (e.g., Web, bases de conhecimento, redes sociais online) a fim de construir coleções locais de dados complexos, representados com base nas estruturas de dados propostas.

Este objetivo tem como foco a construção de amostras locais de redes complexas na Web, a fim de possibilitar a avaliação dos mecanismos para tratamento de informação propostos no âmbito dos demais objetivos específicos relacionados ao desafio 2. Tais amostras deverão ser representativas das redes complexas coletadas e prover exemplos de itens de informação complexos provenientes de diferentes domínios, como bases de conhecimento e redes sociais online.

Atividade 1: Desenvolvimento de políticas de coleta de redes sociais.

Esta atividade contempla o desenvolvimento e avaliação de políticas de seleção para coleta de redes sociais online, a fim de direcionar a obtenção de amostras representativas dessas redes para suporte a tarefas de busca, recomendação, e análise de informação. Além disso, a fim de possibilitar a manutenção de coletas sempre atuais, esta atividade contempla também o desenvolvimento de políticas de revisitação de itens de informação, considerando sua importância nas redes previamente coletadas.

Membros da equipe envolvidos nesta atividade: Nivio Ziviani (UFMG), Jussara Almeida (UFMG), Rodrygo Santos (UFMG), Anderson Ferreira (UFOP), Luiz Henrique Merschmann (UFOP), Evandrino Barros (CEFET), Cristina Murta (CEFET), Humberto Marques Neto (PUCMG), Wladmir Brandão (PUCMG).

Atividade 2: Desenvolvimento de políticas de coleta de dados acadêmicos.

A fim de possibilitar estudos de caso envolvendo redes sociais acadêmicas, esta atividade prevê o desenvolvimento de políticas de seleção para coleta focada de metadados acadêmicos. Tais metadados poderão ser obtidos a partir de páginas pessoais de pesquisadores, seus perfis e de suas publicações em

redes sociais online, máquinas de busca acadêmicas. Finalmente, esta atividade prevê também a coleta de dados de citações, obtidos a partir da coleta em largura do texto completo de publicações.

Membros da equipe envolvidos nesta atividade: Alberto Laender (UFMG), Marcos Gonçalves (UFMG), Rodrygo Santos (UFMG), Anderson Ferreira (UFOP), Luiz Henrique Merschmann (UFOP), Evandrino Barros (CEFET), Cristina Murta (CEFET), Humberto Marques Neto (PUCMG), Wladmir Brandão (PUCMG).

7.2.3 Objetivo 3: Desenvolver mecanismos para a integração de dados provenientes de múltiplas fontes heterogêneas, de modo a unificar suas diversas e potencialmente conflitantes versões em itens de informação consolidados.

Este objetivo contempla o desenvolvimento de mecanismos para promoção e manutenção da integridade dos dados utilizados na execução das demais atividades no âmbito do desafio 2, dados esses obtidos como parte das atividades do Objetivo 2. Em particular, a integridade dos dados poderá ser promovida por meio da integração e reconciliação de múltiplas versões potencialmente discrepantes desses dados, obtidas a partir de fontes de dados heterogêneas com diferentes níveis de autoridade.

Atividade 1: Desenvolvimento de mecanismos para integração de dados.

Esta atividade envolve o desenvolvimento de mecanismos para integração de dados obtidos a partir de múltiplas fontes heterogêneas. Em particular, deverão ser desenvolvidos mecanismos para desambiguação de itens de informação complexos e de seus relacionamentos heterogêneos, considerando-se a similaridade entre as redes egocêntricas induzidas por cada par de itens analisados. Esta atividade prevê também o desenvolvimento de mecanismos para resolução de eventuais conflitos resultantes do processo de integração, incluindo estratégias para versionamento de dados.

Membros da equipe envolvidos nesta atividade: Alberto Laender (UFMG), Marcos Gonçalves (UFMG), Clodoveu Davis (UFMG), Anderson Ferreira (UFOP), Rodrygo Santos (UFMG), Evandrino Barros (CEFET), Cristina Murta (CEFET), Humberto Marques Neto (PUCMG), Wladmir Brandão (PUCMG).

Atividade 2: Desenvolvimento de modelos de qualidade de dados.

A fim de prover uma estimativa quantificável da qualidade dos dados coletados e integrados como resultado da atividade anterior, esta atividade envolve o desenvolvimento de modelos de qualidade de dados. Os modelos propostos devem levar em consideração os níveis de completude e correção de cada item de informação em uma coleção, o nível de consenso existente entre múltiplas versões de um mesmo item, e o nível de autoridade das fontes de dados a partir das quais cada versão foi derivada.

Membros da equipe envolvidos nesta atividade: Wagner Meira Jr. (UFMG), Mirella Moro (UFMG), Alberto Laender (UFMG), Rodrygo Santos (UFMG).

7.2.4 Objetivo 4: Desenvolver mecanismos para enriquecimento dos itens de informação consolidados, incluindo abordagens para reconhecimento de entidades, ligação semântica a bases de conhecimento, anotação, e classificação automática.

Este objetivo contempla o enriquecimento de itens de informação coletados e consolidados em uma coleção. Em particular, são previstas atividades de pesquisa envolvendo o enriquecimento do conteúdo textual de cada item e sua ligação semântica a bases de conhecimento externas.

Atividade 1: Desenvolvimento de mecanismos para enriquecimento textual.

Esta atividade contempla o desenvolvimento de mecanismos para análise de sentimentos, anotação, e classificação automática de itens de informação. Tais mecanismos servirão como forma de aprimorar a organização de coleções de itens de informação à luz de taxonomias predefinidas ou inferidas. Além disso, a descoberta de propriedades latentes de cada item deverão possibilitar relacioná-los a outros itens.

Membros da equipe envolvidos nesta atividade: Wagner Meira Jr. (UFMG), Adriano Veloso (UFMG), Gisele Pappa (UFMG), Loïc Cerf (UFMG), Jussara Almeida (UFMG), Marcos Gonçalves (UFMG), Clodoveu Davis (UFMG) .

Atividade 2: Desenvolvimento de mecanismos para enriquecimento semântico.

De modo complementar à atividade anterior, esta atividade prevê o desenvolvimento de mecanismos para reconhecimento de entidades mencionadas no conteúdo textual de itens de informação. Em contrapartida à simples anotação desse conteúdo, esta atividade prevê também a ligação de itens de informação a bases de conhecimento externas, como a Wikipédia. Tais ligações proporcionarão o enriquecimento semântico de cada item e possibilidades adicionais para o desenvolvimento de mecanismos de busca, recomendação, e análise, conforme propostos a seguir.

Membros da equipe envolvidos nesta atividade: Adriano Veloso (UFMG), Gisele Pappa (UFMG), Alberto Laender (UFMG), Clodoveu Davis (UFMG), Anderson Ferreira (UFOP), Luiz Henrique Merschmann (UFOP), Humberto Marques Neto (PUCMG), Wladmir Brandão (PUCMG).

7.2.5 Objetivo 5: Desenvolver modelos e algoritmos de recuperação de informação, mineração de dados e aprendizado de máquina para tarefas como inferir a relevância de itens de informação complexos dada uma necessidade de informação complexa, de modo a suportar tarefas analíticas, assim como tarefas de busca e recomendação.

Com base nos dados coletados, consolidados, e enriquecidos como resultado das atividades elencadas no âmbito do desafio 2, este objetivo contempla o desenvolvimento de mecanismos para descoberta e recuperação de informação em redes complexas. Tais mecanismos incluem tarefas analíticas, como a mineração de padrões relevantes, bem como tarefas interativas de busca e recomendação.

Atividade 1: Desenvolvimento de mecanismos para busca em redes semânticas.

Esta atividade contempla o desenvolvimento de modelos de entendimento de consultas e de

ranqueamento em redes semânticas, como redes sociais online. Modelos de entendimento de consultas deverão permitir o reconhecimento de menções a entidades indexadas nas consultas submetidas por um usuário. Por sua vez, modelos de ranqueamento deverão considerar tanto a existência de evidências textuais como a de relacionamentos heterogêneos entre necessidades e itens de informação.

Membros da equipe envolvidos nesta atividade: Rodrygo Santos (UFMG), Alberto Laender (UFMG), Nivio Ziviani (UFMG)), Wladmir Brandão (PUCMG).

Atividade 2: Desenvolvimento de mecanismos para recomendação em dispositivos móveis.

De modo complementar à atividade anterior, esta atividade prevê o desenvolvimento de modelos para recomendação contextual em dispositivos móveis. Tais modelos terão como foco a exploração do contexto imediato geográfico e temporal do usuário e dos itens recomendáveis. Além disso, esta atividade prevê o desenvolvimento de modelos de inferência da premência de uma recomendação, de modo a oferecer recomendações relevantes sem a necessidade de uma solicitação explícita do usuário.

Membros da equipe envolvidos nesta atividade: Rodrygo Santos (UFMG), Nivio Ziviani (UFMG), Clodoveu Davis (UFMG).

Atividade 3: Desenvolvimento de mecanismos para descoberta de padrões em redes acadêmicas.

Esta atividade contempla o desenvolvimento de ferramentas para descoberta, análise, e visualização de padrões relevantes em redes sociais acadêmicas. Padrões de interesse incluem, dentre outros, indicadores de produtividade, tópicos de especialidade, e perfis migratórios, segmentados para diferentes estratos demográficos e de áreas do conhecimento.

Membros da equipe envolvidos nesta atividade: Alberto Laender (UFMG), Wagner Meira Jr. (UFMG), Mirella Moro (UFMG), Clodoveu Davis (UFMG), Rodrygo Santos (UFMG).

Atividade 4: Desenvolvimento de mecanismos para descoberta de padrões em fluxos de dados.

Desenvolver mecanismos para mineração de padrões relevantes em fluxos de dados, com vistas à identificação de tendências bem como a detecção de eventos emergentes.

Membros da equipe envolvidos nesta atividade: Wagner Meira Jr. (UFMG), Rodrigo dos Santos (UFJF), Alex Vieira (UFJF), Leonardo Rocha (UFSJ), Humberto Marques Neto (PUCMG).

7.2.6 Cronograma de Execução das Atividades do Desafio 2

O cronograma na Tabela 3 organiza as atividades descritas na Seção 7.2 temporalmente, ao longo dos 48 meses de duração desta proposta.

7.3 Principais Atividades do Desafio 3

O desafio 3 consiste no desenvolvimento de novos algoritmos, infra-estruturas, plataformas e interfaces potencializando implementação de novas aplicações Web de aplicações com alto desempenho. Em

geral, os objetivos específicos deste desafio descritos na Seção 4 visam abstrair e facilitar acesso à funcionalidades do *hardware*. Abaixo descrevemos as atividades previstas para alcançar estes objetivos.

7.3.1 Objetivo 1: Monitorar, caracterizar e modelar a infra-estrutura de redes sobre a qual se apoia a Web e as redes sociais, tanto no nível físico quanto de componentes de software (como topologia e capacidade de enlaces físicos ou plataformas de distribuição de conteúdo) a fim de se entender seu impacto sobre o desempenho das aplicações.

Este objetivo visa melhor compreender o contexto onde aplicações Web são executadas, em particular das diferentes camadas de rede e de software. Apesar de anos de pesquisa desenvolvendo técnicas para monitorar desempenho de redes, técnicas existentes nem sempre são satisfatórias e podem ser inadequadas para estudar aplicações Web. Pretendemos desenvolver novas técnicas de monitoramento, realizar caracterizações e criar modelos analíticos do comportamento da infra-estrutura e seu impacto em aplicações Web.

Atividade 1. Extensão e desenvolvimento de técnicas de monitoramento de rede e de centros de processamento.

Esta atividade visa desenvolver novas técnicas de monitoramento e melhorar técnicas existentes de forma a capturar métricas de desempenho que impactam desempenho de aplicações Web de forma precisa e com baixa sobrecarga. Consideramos captura de dados de forma passiva (e.g., monitorando tráfego de rede existente e carga de aplicações Web [81, 101, 106]) e ativa (e.g., enviando sondas na rede e fazendo requisições a aplicações [38, 57, 102]). O resultado esperado é coletar uma base de dados relacionando medições precisas de várias métricas de desempenho em diferentes cenários com o desempenho de aplicações Web.

Membros da equipe envolvidos nesta atividade: Cristina Duarte Murta (CEFET-MG), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

Atividade 2. Caracterização de dados de rede, centros de processamento de dados e aplicações.

Iremos implantar as técnicas de monitoramento desenvolvidas na atividade 1 em redes e centros de processamento de dados reais para coletar informações. Iremos caracterizar estes dados para melhor entender o comportamento e desempenho de infra-estruturas em diferentes cenários como falhas de hardware, congestionamento de rede, sobrecarga de trabalho e *flash crowds*. Além de entender melhor infra-estruturas de rede, iremos relacionar o comportamento e o desempenho de infra-estruturas com o desempenho de aplicações Web.

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Cristina Duarte Murta (CEFET-MG), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

Atividade 3. Modelagem do impacto de anomalias no desempenho de aplicações Web.

Utilizaremos o melhor entendimento obtido das caracterizações realizadas na atividade 2 para propor modelos que generalizem os resultados da caracterização. Nosso objetivo é propor modelos que permitam análises do tipo "o que aconteceria se" em diversos cenários. Tais modelos podem ser utilizados para tomada de decisões, previsões, provisionamento e como etapa preliminar de avaliação de desempenho.

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Cristina Duarte Murta (CEFET-MG), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

7.3.2 Objetivo 2: Explorar infra-estruturas de redes (overlays, caches, e centros de processamento na nuvem) para serviços baseados em redes sociais, máquinas de busca e distribuição de conteúdo.

Existem diversas tecnologias, mecanismos e plataformas para implementação de infra-estruturas de rede e desenvolvimento de serviços Web (e.g., [9, 16, 33, 43, 59, 93]). Este objetivo consiste em estudar de forma sistemática destas infra-estruturas e criar processos para permitir a escolha da infra-estrutura mais adequada para um serviço Web. Orientadas pelos resultados do objetivo 1, as primeiras duas atividades relacionam características de infra-estruturas de redes com requisitos de serviços Web. As demais atividades aplicam os resultados destas duas etapas para criação de um processo decisório para escolha de infra-estruturas e finalmente avaliação de serviços Web resultantes.

Atividade 1. Identificação e modelagem de infra-estruturas distribuídas.

Iremos estudar mecanismos para implementação de infra-estruturas (como DHTs [92, 117], técnicas de replicação [34], códigos de correção de erro [95], virtualização [16, 59], algoritmos distribuídos e paralelos [65, 111]) e arquiteturas distribuídas existentes (como overlays [9], nuvens [1, 2, 93], P2P [33]), analisando os compromissos e funcionalidades de cada mecanismo e sistema (como escalabilidade, sobrecarga e robustez).

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

Atividade 2. Mapeamento de requisitos de serviços Web.

Iremos mapear os requisitos de serviços Web (como redes sociais, máquinas de busca e distribuição de conteúdo) relacionando-os com mecanismos e arquitetura de infra-estrutura considerados na atividade 1. Esta tarefa é complicada por diferenças que fazem aplicações Web mais ou menos suscetíveis a problemas de desempenho distintos bem como dificuldades em quantificar o impacto de problemas de desempenho na percepção do usuário final [56]. Nosso objetivo é identificar quais mecanismos e sistemas são mais adequados para implementação de cada tipo de serviço Web.

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Fabrício Benevenuto (UFMG), Ítalo Cunha (UFMG), Jussara Almeida (UFMG), Rodrigo Weber dos Santos (UFJF).

Atividade 3. Desenvolvimento de arcabouço para prototipagem de serviços Web.

Iremos desenvolver um arcabouço para implementação de serviços Web em *gateways* domésticos inteligentes [43, 62]. Nossa plataforma prevê criação e gerenciamento de máquinas virtuais que poderão ser utilizadas para prototipagem e instalação de serviços Web em infra-estruturas alternativas reais. O objetivo 3 (seção 7.3.3) complementa esta atividade e prevê desenvolvimento de arcabouço para prototipagem de serviços Web em infra-estruturas de alta capacidade.

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

Atividade 4. Implementação e avaliação de serviços Web descentralizados.

Iremos utilizar o mapeamento de requisitos realizado nas atividades 1 e 2 para guiar a implementação de serviços Web descentralizados. Focaremos na implementação de protótipos para avaliação mais realista das infra-estruturas propostas, com possível avaliação complementar via simulações se necessário (e.g., para análise de escalabilidade). Avaliaremos métricas como escalabilidade, sobrecarga, mantenabilidade e robustez dos serviços implementados.

Membros da equipe envolvidos nesta atividade: Alex B. Vieira (UFJF), Dorgival Guedes (UFMG), Evandrino G. Barros (CEFET-MG), Ítalo Cunha (UFMG), Rodrigo Weber dos Santos (UFJF).

7.3.3 Objetivo 3: Desenvolver ambientes para implementação, suporte em tempo de execução e análise de desempenho para algoritmos paralelos e distribuídos em arquiteturas multi-core, many-core e heterogêneas multi-nível.

Serviços Web frequentemente ultrapassam a faixa de milhões de usuários e impões requisitos estritos de desempenho, dependabilidade, escalabilidade e gerência sobre a infra-estrutura subjacente. Neste objetivo iremos desenvolver e implementar ambientes para implementação de serviços Web em arquiteturas de hardware paralelas e distribuídas.

Atividade 1. Escalonamento em ambientes heterogêneos.

As propostas desse projeto demandam a execução de um grande número de tarefas com perfis bastante diferenciados, em hardware também bastante diferenciado. É necessário, com isso, construir mecanismos que permitem a alocação das tarefas aos diversos processadores do ambiente de execução.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Renato Ferreira (UFMG).

Atividade 2. Reconfiguração dinâmica.

Os cenários de aplicação propostos são de execuções de longa duração. É então necessário a construção de mecanismos de rebalanceamento dinâmico da carga, para responder às necessidades que variam ao longo do tempo entre as diversas tarefas das aplicações.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Renato Ferreira (UFMG).

Atividade 3. Programação em alto nível.

As aplicações desenvolvidas são de alta complexidade, e a plataforma de execução é também sofisticada devido aos requisitos impostos pelo projeto. Programar as aplicações, assim, é uma tarefa completa. Nesse sentido, pretendemos desenvolver ferramentas de suporte à programação, que permita aos desenvolvedores escrever suas aplicações de maneira natural ao mesmo tempo que produz um código capaz de executar eficiente no hardware disponível.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Fernando M. Q. Pereira (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Renato Ferreira (UFMG).

Atividade 4. Suporte a profiling.

Pretendemos desenvolver uma console interativa onde o administrador do sistema consegue enxergar a execução das diversas aplicações no hardware distribuído. Essa ferramenta será útil não somente no contexto de uma aplicação específica, ao facilitar a detecção de eventuais gargalos, mas também é útil ao sistema, pois permite reconhecer gargalos constituídos pelo coletivo das aplicações que compartilham o hardware.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Fernando M. Q. Pereira (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Renato Ferreira (UFMG).

7.3.4 Objetivo 4: Paralelizar algoritmos de recuperação de informação, mineração de dados e aprendizado de máquina para arquiteturas paralelas e distribuídas.

Além de melhorias na infra-estrutura utilizada para implantação de serviços Web descritas no desafio anterior, iremos também projetar novos algoritmos paralelos para implementação de serviços Web de maior desempenho e com custo (de operação) reduzido.

Atividade 1. Seleção de algoritmos a serem paralelizados.

Dentre os algoritmos gerados ou utilizados no contexto do projeto, identificar aqueles que sejam demandantes em termos de recursos computacionais e possuam oportunidades de paralelização.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Humberto Torres Marques Neto (PUC-MG), Renato Ferreira (UFMG), Wagner Rodrigo Weber dos Santos (UFJF), Meira Jr. (UFMG), Wladmir Cardoso Brandão (PUC-MG).

Atividade 2. Apropriação dos algoritmos.

Nesta atividade iremos analisar detalhadamente cada algoritmo selecionado, verificando as peculiaridades da implementação sequencial do mesmo e identificando as possíveis estratégias de paralelização a serem utilizadas.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Humberto Torres Marques Neto (PUC-MG), Renato Ferreira (UFMG), Wagner

Rodrigo Weber dos Santos (UFJF), Meira Jr. (UFMG), Wladmir Cardoso Brandão (PUC-MG).

Atividade 3. Projeto dos algoritmos paralelos.

Tendo em vista o ambiente de programação a ser utilizado e as características do modelo de programação, as versões paralelas dos algoritmos são projetadas, levando em consideração aspectos como localidade de referência, sobreposição de computação e comunicação e balanceamento de carga.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Humberto Torres Marques Neto (PUC-MG), Rodrigo Weber dos Santos (UFJF), Renato Ferreira (UFMG), Wagner Meira Jr. (UFMG), Wladmir Cardoso Brandão (PUC-MG).

Atividade 4. Implementação dos algoritmo paralelos.

Os algoritmos paralelos serão implementados e, dentro das suas oportunidades de paralelismo, podem explorar três tipos de paralelismo: assincronia, tarefas e dados. Eventualmente o algoritmo pode ser implementado para mais de um ambiente de execução, conforme necessidade.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Humberto Torres Marques Neto (PUC-MG), Rodrigo Weber dos Santos (UFJF), Renato Ferreira (UFMG), Wagner Meira Jr. (UFMG), Wladmir Cardoso Brandão (PUC-MG).

Atividade 5. Validação dos algoritmos.

Os algoritmos serão validados experimentalmente, variando parâmetros de execução tanto em termos da aplicação que foi paralelizada, quanto em termos do ambiente de execução, como número de processadores e natureza dos processadores.

Membros da equipe envolvidos nesta atividade: Dorgival Guedes (UFMG), Leonardo Chaves Dutra da Rocha (UFSJ), Humberto Torres Marques Neto (PUC-MG), Rodrigo Weber dos Santos (UFJF), Renato Ferreira (UFMG), Wagner Meira Jr. (UFMG), Wladmir Cardoso Brandão (PUC-MG).

7.3.5 Objetivo 5: Investigar estratégias de interação que permitam aos usuários um melhor aproveitamento de serviços da Web no seu contexto, tais como personalização, programação pelo usuário final e visualização.

Para que a entrega da informação possa ser considerada satisfatória, a interface deve ser capaz de apresentá-la de forma que o usuário a entenda. Além de entendê-la para que o usuário possa fazer melhor uso da informação no seu contexto, ele deve ser capaz de interagir com ela, definindo que aspectos da informação deseja receber e como.

Atividade 1. Investigação de como e em que contextos algoritmos de análise de sentimentos podem apoiar usuários na seleção da informação desejada.

Atualmente a grande quantidade de informações pode requerer diferentes tipos de filtros para facilitar a seleção dos usuários sobre o tipo de informação que ele pretende acessar. Algoritmos de análise de sentimento permitem que se avalie o sentimento associado à informação como positiva ou negativa.

Pretende-se investigar como e em que situações algoritmos de análise de sentimento podem apoiar o usuário na seleção da informação que ele pretende acessar.

Membros da equipe envolvidos nesta atividade: Fabrício Benevenuto (UFMG), Raquel Prates (UFMG).

Atividade 2. Uso de simulação na antecipação dos resultados de configurações possíveis

Diversos sistemas colaborativos na Web 2.0 permitem que usuários configurem por quem ou em que situações uma informação pode ser acessada. A combinação das diversas configurações feitas pelos usuários e ações resultantes possibilitam a definição de diferentes processos sociais que podem ser vivenciados pelos usuários. Será explorada como a simulação pode apoiar o usuário na antecipação destes diferentes processos sociais e seus possíveis impactos.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

Atividade 3. Análise da aplicabilidade de uma ontologia para recomendação de visualizações para dados qualitativos

A Web 2.0 com frequencia disponibiliza grandes volumes de dados a serem apresentados pelos usuários. Já se tem disponível ontologias que descrevem as relações quantitativas sendo descritas, os padrões visuais revelados, assim como, as técnicas de navegação e interação analíticas que possam ser usadas na visualização. Nesta atividade, pretende-se analisar se e como essas ontologias podem ser úteis também na visualização de dados qualitativos.

Membros da equipe envolvidos nesta atividade: Raquel Minardi (UFMG) e Raquel Prates (UFMG) .

Atividade 4. Investigação de qualidade de uso e técnicas de interação

O grande volume de informação e a possibilidade de interagir com diversas pessoas em diferentes locais e pertencentes a culturas distintas e com necessidades diversas traz grandes desafios. Vamos investigar modelos, métodos e técnicas que permitam uma interação de qualidade com grandes volumes de dados, por pessoas que tenham diferentes experiências ou necessidades.

Membros da equipe envolvidos nesta atividade: Raquel Prates (UFMG).

7.3.6 Cronograma de Execução das Atividades do Desafio 3

A seguir apresentamos o cronograma das atividades do desafio 3, descritas nas seções 7.3.1–7.3.5.

8 Recursos Solicitados

Os recursos solicitados neste projeto consideram a sua execução por 4 anos e o seu forte caráter experimental e baseado em implementações dos modelos, algoritmos e sistemas propostos nos vários cenários de aplicação. A tabela a seguir apresenta uma síntese dos recursos solicitados, os quais são detalhados nas seções a seguir.

Rubrica	Tipo de dispêndio	Valor (R$)
Capital	Material permanente e equipamento	75.350,00
Custeio	Material de consumo	37.000,00
	Diárias	175.384,00
	Passagens	216.880,00
Bolsas	Bolsa IC	57.600,00
	Bolsa DTI	219.067,20
	Despesas operacionais	39.064,06
Total		820.345,24

8.1 Capital

Solicitamos capital para equipar os laboratórios das instituições parceiras, o que compreende tanto servidores quanto estações de trabalho e outros equipamentos, como no-breaks e impressora.

8.2 Custeio

Os itens de custeio compreende material de consumo e principalmente diárias e passagens para visitas técnicas e apresentação dos resultados do projeto de pesquisa em fóruns qualificados.

8.3 Bolsas

Solicitamos 3 bolsas na modalidade DTI e três bolsas na modalidade IC. Estes bolsisas irão atuar primariamente no desenvolvimento das bibliotecas e protótipos do projeto.

Dispêndio	Valor Unit	UFMG	PUCMG	CEFET	UFOP	UFJF	UFSJ	Total
Diárias Int	506,00	97.152,00	7.084,00	10.120,00	8.096,00	10.120,00	3.542,00	136.114,00
Passagem Int	4.140,00	99.360,00	16.560,00	16.560,00	16.560,00	16.560,00	8.280,00	173.880,00
Diárias Nac	210,00	20.160,00	4.200,00	2.100,00	2.520,00	8.400,00	1.890,00	39.270,00
Passagem Nac	1.000,00	24.000,00	5.000,00	4.000,00	4.000,00	4.000,00	2.000,00	43.000,00
BDTI-3	1.521,30	219.067,20	0,00	0,00	0,00	0,00	0,00	219.067,20
Iniciação Científica	400,00	57.600,00	0,00	0,00	0,00	0,00	0,00	57.600,00
Custeio	1.000,00	10.000,00	0,00	20.000,00	2.000,00	4.000,00	1.000,00	37.000,00
Capital	1.000,00	0,00	24.000,00	4.000,00	23.050,00	12.600,00	11.700,00	75.350,00
Sub-total								781.281,20
Despesas Operacionais								39.064,06

9 Instalações físicas e equipamentos

A infraestrutura que apoia o MASWeb é constituída de laboratórios de pesquisa bem configurados e com equipamentos especializados de última geração. Uma eficiente equipe de gerenciamento de recursos computacionais dará suporte a todas as atividades do Núcleo. A relação de cooperação com o Departamento de Ciência da Computação, que vem desde projetos anteriores do mesmo grupo, é

de fundamental importância para o seu sucesso. O apoio administrativo oferecido para os discentes e docentes é um dos fatores chave para o bom andamento do Núcleo.

O Núcleo compartilha excelente infraestrutura com o DCC, que inclui instalações em três andares do prédio do Instituto de Ciências Exatas (ICEx). Todos os pesquisadores têm acesso à rede sem fio do DCC, à rede com fio Gigabit Ethernet, à rede sem fio da UFMG, à rede internacional "eduroam". Essa infraestrutura está disponível na modalidade 24x7.

O Núcleo possui vários laboratórios de pesquisa que apoiarão as atividades de pesquisa, conforme mostrado na Tabela 5. Além da infraestrutura de salas e laboratórios, o Núcleo terá acesso a salas de seminários e discussões equipadas com computador, projetor de alta resolução, quadros brancos, aparelho de TV de plasma/led, lousa eletrônica, climatização silenciosa, equipamento de som de alta qualidade e equipamento de vídeo-conferência, o que tem possibilitado a participação remota sem perda de qualidade de professores, pesquisadores e alunos em reuniões e palestras.

O Ponto de Presença da RNP em Minas Gerais (POP-MG), que é quem gerencia todas as conexões de instituições conectadas à RNP em Minas Gerais, inclusive a da UFMG, é operado pelo DCC. Além da garantia de acesso à Internet com qualidade, esse arranjo possibilita um grande laboratório real de redes de computadores.

10 Contrapartida

A contrapartida para a realização do projeto do Núcleo de Excelência para a Web é a seguinte:

- Cluster com 16 servidores de alto desempenho, quad processado, 966Mb RAM e 96Tb de disco no valor total de $320,000.00$ reais.
- 6 Laboratórios que possuem aproximadamente 60 postos de trabalho, todos contando com estações de trabalho atualizadas, no valor de $5,000.00$ reais, totalizando $300,000.00$ reais.

11 Relevância

11.1 Formação de Recursos Humanos

Considerando uma estimativa conservadora, o MASWeb pretende formar anualmente 5 alunos de doutorado, 20 alunos de mestrado e orientar vários alunos em programas de Iniciação Científica. O número atual de alunos em curso nos vários níveis supera essas expectativas.

Um mecanismo particularmente interessante que será adotado na formação de alunos é a realização de visitas técnicas e estágios em instituições no exterior, além da promoção da mobilidade de alunos entre as instituições que compoem o MASWeb, reforçando os laços de trabalho nas várias linhas de pesquisa.

11.2 Transferência de Conhecimento e Tecnologia

O grupo que compõe o Núcleo de Excelência para a Web possui vasta experiência no que tange a transferência de conhecimento e tecnologia para o **setor empresarial ou governamental**. Esta experiência é decorrente, principalmente, de resultados relacionados ao Instituto Nacional de Ciência e Tecnologia para a Web (InWeb).

Considerando a transferência de conhecimento para a sociedade (através do governo federal), podemos citar a produção de protótipos que poderão, por exemplo, ser usados pelos setores do poder público. contexto, um protótipo que já se transformou em artefato operacional do Ministério da Saúde é o Observatório da Dengue, que vem sendo utilizado para vigilância epidemiológica de rumores da epidemia da dengue no Brasil e deve ser estendido para outras doenças.

Além da publicação de protótipos, a transferência de conhecimento e tecnologia gerados pelo MASWeb também será feita por meio da criação de start-ups intensivas em conhecimento, i.e., empresas cuja principal fonte de valor advém do conhecimento gerado. O grupo de pesquisadores do Núcleo de Excelência para a Web possui um vasta experiência na criação de start-ups. Um exemplo de sucesso de empresas com essas características é a Zunnit.

A Zunnit é uma empresa especializada em desenvolver sistemas de recomendação, um dos objetivos da linha de Recuperação de Informação. Um sistema de recomendação identifica automaticamente os interesses e o perfil do usuário a partir de seu padrão de comportamento e contexto de navegação e recomenda conteúdo personalizado, como informações, serviços e produtos. A empresa foi criada em 2009 a partir da associação entre os professores Alberto Laender (DCC/UFMG), Altigran Silva (DCC/UFAM), Edleno Moura (DCC/UFAM) e Nivio Ziviani (DCC/UFMG), Alan Castro (Mestre DCC/UFMG), Anísio Lacerda (Doutor DCC/UFMG) e Guilherme Menezes (Mestre DCC/UFMG), e os investidores Alberto Colares e Lesley Scarioli Júnior.

Uma iniciativa importante que auxiliará na transferência de tecnologia, no contexto do MASWeb, é o **CTWeb - Centro Tecnológico para a Web**, cujas atividades se inicial em março de 2014, com sede no BH-TEC – Parque Tecnológico de Belo Horizonte. O CTWeb tem por objetivo ser responsável pela operação de protótipos, maturação de tecnologias geradas no âmbito do InWeb e estabelecimento de convênios e contratos para utilização e transferência de tecnologias. O CTWeb inaugura um modelo inovador de transferência de conhecimento e tecnologia, em particular no contexto de IFES (instituições federais de ensino superior). Por esse modelo, a IFE, no caso a UFMG, licencia o conhecimento e tecnologia advindo da pesquisa para o CTWeb, que pode estendê-lo, sub-licenciá-lo ou estabelecer parcerias. No momento estão sendo negociadas parcerias com três empresa envolvendo mais de 20 tecnologias, todas oriundas do Observatório da Web. Em termos de ressarcimento, há também opções inovadoras, como a cessão de ações ou o compartilhamento de receita líquida. A nossa expectativa é que o CTWeb esteja completamente operacional até o fim de 2014.

11.3 Educação e Divulgação da Ciência

O MASWeb irá promover workshops anuais com participação de todos os membros da equipe, seus alunos e outros membros da comunidade científica, para discussão e avaliação dos principais resultados obtidos bem como para definição de possíveis ajustes e novos direcionamentos.

O pesquisadores do MASWeb já organizaram e participaram da organização de vários eventos nos últimos anos, com destaque para a World Wide Web Conference 2013, em parceria com o INCT de WebScience, sediado na PUC-Rio. Essa é a maior conferência da área de Web e aconteceu no Rio de Janeiro em maio de 2013. O professor Virgilio Almeida foi um dos coordenadores gerais da WWW 2013. Em 2014, já vamos apoiar a realização do 9th Latin American Web Congress, em outubro, em Ouro Preto, MG.

Referências

[1] Amazon Web Services. `http://aws.amazon.com`.

[2] Windows Azure. `https://www.windowsazure.com`.

[3] Mark S. Ackerman. The intellectual challenge of cscw: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2-3):179–203, 2000.

[4] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 607–616, New York, NY, USA, 2013. ACM.

[5] Miltiadis Allamanis, Salvatore Scellato, and Cecilia Mascolo. Evolution of a location-based online social network: Analysis and models. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pages 145–158, New York, NY, USA, 2012. ACM.

[6] Jussara Almeida, Marcos Gonçalves, and Raquel Prates. Towards the green web: Fighting pollution and promoting high quality content on the web. In *Proceedings of the 3rd International ACM Conference on Web Science*, WebSci'11, pages 1–3, 2011.

[7] Jussara M. Almeida, Virgílio A. F. Almeida, Danilo Ardagna, Chiara Francalanci, and Marco Trubian. Resource management in the autonomic service-oriented architecture. In *ICAC*, pages 84–92. IEEE, 2006.

[8] Sean Amiratti. Google's udi manber- search is a hard problem, June 2007.

[9] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient Overlay Networks. *SIGOPS Oper. Syst. Rev.*, 35(5):131–145, 2001.

[10] Stephanos Androutsellis-Theotokis and Diomidis Spinellis. A survey of peer-to-peer content distribution technologies. *ACM Comput. Surv.*, 36(4):335–371, 2004.

[11] Pedro Antunes, Valeria Herskovic, Sergio F. Ochoa, and Jose A. Pino. Structuring dimensions for collaborative systems evaluation. *ACM Comput. Surv.*, 44(2):8:1–8:28, March 2008.

[12] Algirdas Avizienis, Jean-Claude Laprie, Brian Randell, and Carl E. Landwehr. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable Sec. Comput.*, 1(1):11–33, 2004.

[13] Robert Axelrod. *The Evolution of Cooperation: Basic Books*. Basic Books, 1984.

[14] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 2009.

[15] Luciano Barbosa, Juliana Freire, and Altigran Soares da Silva. Organizing hidden-web databases by clustering visible web documents. In Rada Chirkova, Asuman Dogac, M. Tamer Özsu, and Timos K. Sellis, editors, *ICDE*, pages 326–335. IEEE, 2007.

[16] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the Art of Virtualization. In *Proc. ACM SOSP*, 2003.

[17] Louise Barkhuus, Barry Brown, Marek Bell, Scott Sherwood, Malcolm Hall, and Matthew Chalmers. From awareness to repartee: sharing location within social groups. In Mary Czerwinski, Arnold M. Lund, and Desney S. Tan, editors, *CHI*, pages 497–506. ACM, 2008.

[18] Fabiano Belém, Eder Martins, Tatiana Pontes, Jussara Almeida, and Marcos Gonçalves. Associative tag recommendation exploiting multiple textual features. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1033–1042, New York, NY, USA, 2011. ACM.

[19] Fabiano Belém, Rodrygo Santos, Jussara Almeida, and Marcos Gonçalves. Topic diversity in tag recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 141–148, New York, NY, USA, 2013. ACM.

[20] Fabiano Muniz Belém, Eder Ferreira Martins, Jussara Marques Almeida, and Marcos André Gonçalves. Exploiting relevance, novelty and diversity in tag recommendation. In *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web*, WebMedia '12, pages 297–300, New York, NY, USA, 2012. ACM.

[21] F. Benevenuto, T. Rodrigues, A. Veloso, J Almeida, M. Goncalves, and V. Almeida. Practical detection of spammers and content promoters in online video sharing systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(3):688–701, June 2012.

[22] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 620–627, New York, NY, USA, 2009. ACM.

[23] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, Marcos Gonçalves, and Keith Ross. Video pollution on the web. *First Monday*, 15:4–5, 2010.

[24] Saul J. Berman, Bill Battino, Louisa Shipnuck, and Andreas Neus. The end of advertising as we know it. Technical report, IBM Institute for Business Value, 2007.

[25] Michele A. Brandão, Mirella M. Moro, Giseli Rabello Lopes, and José Palazzo Moreira de Oliveira. Using link semantics to recommend collaborations in academic social networks. In *WWW (Companion Volume)*, pages 833–840, 2013.

[26] John Seely Brown and Paul Duguid. *The Social Life of Information*. Harvard Business Press, 2000.

[27] Moira Burke and Robert Kraut. Mind your p's and q's: When politeness helps and hurts in online communities. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3195–3200, New York, NY, USA, 2008. ACM.

[28] Pável Calado, Marco Cristo, Marcos André Gonçalves, Edleno S. de Moura, Berthier Ribeiro-Neto, and Nivio Ziviani. Linkage similarity measures for the classification of web

documents. *Journal of the American Society for Information Science and Technology*, 57(2):208–221, 2006.

[29] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Network*, 17(5):1357–1370, October 2009.

[30] F Chang, J Dean, S Ghemawat, WC Hsieh, DA Wallach, M Burrows, T Chandra, A Fikes, and RE Gruber. Bigtable: A distributed storage system for structured data. *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI'06)*, 2006.

[31] Chaomei Chen. Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, 25(4):12–16, 2005.

[32] Dave Clark, Bill Lehr, Steve Bauer, Peyman Faratin, Rahul Sami, and John Wroclawski. Overlay Networks and the Future of the Internet. *COMMUNICATIONS & STRATEGIES*, (63), 2006.

[33] B. Cohen. Incentives Build Robustness in BitTorrent. In *Proc. Workshop on Economics of Peer-to-Peer Systems*, 2003.

[34] E. Cohen and S. Shenker. Replication Strategies in Unstructured Peer-to-peer Networks. In *Proc. ACM SIGCOMM*, 2002.

[35] Cristiano P. Costa, Ítalo S. Cunha, Alex Borges Vieira, Claudiney Vander Ramos, Marcus V. M. Rocha, Jussara M. Almeida, and Berthier A. Ribeiro-Neto. Analyzing client interactivity in streaming media. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *WWW*, pages 534–543. ACM, 2004.

[36] Thierson Couto, Marco Cristo, Marcos André Gonçalves, Pável Calado, Nivio Ziviani, Edleno Silva de Moura, and Berthier A. Ribeiro-Neto. A comparative study of citations and links in document classification. In Gary Marchionini, Michael L. Nelson, and Catherine C. Marshall, editors, *JCDL*, pages 75–84. ACM, 2006.

[37] R Crane and D Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[38] Í. Cunha, R. Teixeira, D. Veitch, and C. Diot. Predicting and Tracking Internet Path Changes. In *Proc. ACM SIGCOMM*, 2011.

[39] Ítalo S. Cunha, Jussara M. Almeida, Virgilio Almeida, and Marcos Santos. Self-adaptive capacity management for multi-tier virtualized environments. In *Integrated Network Management*, pages 129–138. IEEE, 2007.

[40] Ítalo S. Cunha, Itamar Viana, João Palotti, Jussara M. Almeida, and Virgilio Almeida. Analyzing security and energy tradeoffs in autonomic capacity management. In *NOMS*, pages 302–309. IEEE, 2008.

[41] Guilherme T. de Assis, Alberto H. F. Laender, Marcos André Gonçalves, and Altigran Soares da Silva. Exploiting genre in focused crawling. In Nivio Ziviani and Ricardo A. Baeza-Yates, editors, *SPIRE*, volume 4726 of *Lecture Notes in Computer Science*, pages 62–73. Springer, 2007.

[42] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150. USENIX Association, 2004.

[43] C. Dixon, R. Mahajan, S. Agarwal, A. J. Brush, B. Lee, S. Saroiu, and P. Bahl. An Operating System for the Home. In *Proc. USENIX NSDI*, 2012.

[44] Flavio Figueiredo, Fabrício Benevenuto, and Jussara Almeida. The tube over time: Characterizing popularity growth of youtube videos. In *Proceedings of the 4th ACM International Conference of Web Search and Data Mining*, 2011.

[45] A. Finamore, M. Mellia, M. Meo, M.M. Munafo, and D. Rossi. Experiences of internet traffic monitoring with tstat. *Network, IEEE*, 25(3):8–14, May 2011.

[46] Qin Gao, Yusen Dai, Zao Fan, and Ruogu Kang. Understanding factors affecting perceived sociability of social software. *Computers in Human Behavior*, 26(6):1846 – 1861, 2010. Online Interactivity: Role of Technology in Behavior Change.

[47] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In Michael L. Scott and Larry L. Peterson, editors, *SOSP*, pages 29–43. ACM, 2003.

[48] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabrício Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 61–70, New York, NY, USA, 2012. ACM.

[49] Luíz Henrique Gomes, Cristiano Cazita, Jussara M. Almeida, Virgílio A. F. Almeida, and Wagner Meira Jr. Workload models of spam and legitimate e-mails. *Perform. Eval.*, 64(7-8):690–714, 2007.

[50] M.A. Gonçalves, J Almeida, L.G.P. dos Santos, A.H.F. Laender, and V. Almeida. On popularity in the blogosphere. *Internet Computing, IEEE*, 14(3):42–49, May 2010.

[51] Alon Y Halevy, Anand Rajaraman, and Joann J Ordille. Data Integration: The Teenage Years. In *VLDB*, pages 9–16, 2006.

[52] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: An interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, July 2008.

[53] James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel Weitzner. Web science: An interdisciplinary approach to understanding the web. *Communication of the ACM*, 51(7):60–69, 2008.

[54] Lucila Ishitani, Virgilio Almeida, and Wagner Meira Jr. Masks: Bringing anonymity and personalization together. *IEEE Security and Privacy*, 1(3):18–23, May 2003.

[55] Judith L. Jerkins and Jonathan L. Wang. From network measurement collection to traffic performance modeling: challenges and lessons learned. *J. Braz. Comp. Soc.*, 5(3), 1999.

[56] D. Joumblatt, R. Teixeira, J. Chandrashekar, and N. Taft. Performance of Networked Applications: The Challenges in Capturing the User's Perception. In *Proc. ACM SIGCOMM Workshop on Measurements Up the Stack*, 2011.

[57] E. Katz-Bassett, C. Scott, D. R. Choffnes, I. Cunha, V. Valancius, N. Feamster, H. V. Madhyastha, T. Anderson, and A. Krishnamurthy. LIFEGUARD: Practical Repair of Persistent Route Failures. In *Proc. ACM SIGCOMM*, 2012.

[58] William Sean Kennedy, Jamie Morgenstern, Gordon Wilfong, and Lisa Zhang. Hierarchical community decomposition via oblivious routing techniques. In *Proceedings of the First ACM Conference on Online Social Networks*, pages 107–118, 2013.

[59] A. Kivity, Y. Kamay, D. Laor, U. Lublin, and A. Liguori. KVM: the Linux Virtual Machine Monitor. In *Proc. of the Linux Symposium*, 2007.

[60] Alberto H. F. Laender, Carlos J. P. de Lucena, José Carlos Maldonado, Edmundo de Souza e Silva, and Nivio Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bull.*, 40(2):135–145, 2008.

[61] Alberto H. F. Laender, Berthier A. Ribeiro-Neto, and Altigran Soares da Silva. Debye - data extraction by example. *Data Knowl. Eng.*, 40(2):121–154, 2002.

[62] N. Laoutaris, P. Rodriguez, and L. Massoulie. ECHOS: Edge Capacity Hosting Overlays of Nano Data Centers. *SIGCOMM Comput. Commun. Rev.*, 38(1):51–54, 2008.

[63] Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In *Proceedings of the 10th International Conference on Pervasive Computing*, Pervasive'12, pages 91–98, Berlin, Heidelberg, 2012. Springer-Verlag.

[64] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863, April 2009.

[65] N. A. Lynch. *Distributed Algorithms*. Morgan Kaufmann Publishers Inc., 1996.

[66] Michelle Madejski, Maritza Lupe Johnson, and Steven Michael Bellovin. The failure of online social network privacy settings. Technical report, Department of Computer Science, Columbia University, 2011.

[67] Ami Marowka. The grid: Blueprint for a new computing infrastructure. *Scalable Computing: Practice and Experience*, 3(3), 2000.

[68] Eder F. Martins, Fabiano M. Belém, Jussara M. Almeida, and Marcos Gonçalves. Measuring and addressing the impact of cold start on associative tag recommenders. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web*, pages 325–332, New York, NY, USA, 2013. ACM.

[69] Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 6–14, New York, NY, USA, 2012. ACM.

[70] Claudia Bauzer Medeiros. Grand research challenges in computer science in brazil. *IEEE Computer*, 41(6):59–65, 2008.

[71] Johnnatan Messias, Lucas Schmidt, Ricardo Oliveira, and Fabrício Benevenuto. You followed my bot! transforming robots into influential users in twitter. *First Monday*, 18(7), 2013.

[72] Elinor Mills. Google still waiting for social ad payoff, January 2008.

[73] Jelena Mirkovic and Peter Reiher. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2):39–53, April 2004.

[74] Aarti Munjal, Tracy Camp, and William C. Navidi. Smooth: A simple way to model human mobility. In *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '11, pages 351–360, New York, NY, USA, 2011. ACM.

[75] Dorgival Olavo Guedes Neto, Wagner Meira Jr., and Renato Ferreira. Anteater: A service-oriented architecture for high-performance data mining. *IEEE Internet Computing*, 10(4):36–43, 2006.

[76] E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[77] S. Nikolov. *Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series*. PhD thesis, Massachusetts Institute of Technology, 2012.

[78] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153, 2012.

[79] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[80] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, volume WS-11-02 of *AAAI Workshops*. AAAI, 2011.

[81] V. Padmanabhan, L. Qiu, and H. Wang. Server-based Inference of Internet Link Lossiness. In *Proc. IEEE INFOCOM*, 2003.

[82] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA, 2004.

[83] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 365–374, New York, NY, USA, 2013. ACM.

[84] Jennifer Preece. *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, September 2000.

[85] Jenny Preece. *Online Communities: Designing Usability and Supporting Socialbilty*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 2000.

[86] Jenny Preece. Etiquette online: From nice to necessary. *Commun. ACM*, 47(4):56–61, April 2004.

[87] Kira Radinsky, Krysta M. Svore, Susan T. Dumais, Milad Shokouhi, Jaime Teevan, Alex Bocharov, and Eric Horvitz. Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Trans. Inf. Syst.*, 31(3):16:1–16:37, August 2013.

[88] B. Randell and J.N. Buxton. *Software Engineering Techniques*. NATO, Scientific Affairs Division, Brussels, 1970.

[89] B.G. Rocha, V. Almeida, and D. Guedes. Increasing qos in selfish overlay networks. *Internet Computing, IEEE*, 10(3):24–31, May 2006.

[90] Bruno Gusmao Rocha, Virgilio Almeida, and Dorgival Guedes. Increasing qos in selfish overlay networks. *IEEE Internet Computing*, 10(3):24–31, May 2006.

[91] Marcus V. M. Rocha, Marcelo Maia, Ítalo S. Cunha, Jussara M. Almeida, and Sérgio Vale Aguiar Campos. Scalable media streaming to interactive users. In HongJiang Zhang, Tat-Seng Chua, Ralf Steinmetz, Mohan S. Kankanhalli, and Lynn Wilcox, editors, *ACM Multimedia*, pages 966–975. ACM, 2005.

[92] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In *Proceedings of the IFIP/ACM International Conference on Distributed Systems*, 2001.

[93] J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratnasamy, and V. Sekar. Making Middleboxes Someone else's Problem: Network Processing As a Cloud Service. In *Proc. ACM SIGCOMM*, 2012.

[94] B. Shneiderman and C. Plaisant. *Information Search and Visualization*. Designing the User Interface. Addison Wesley, 4 edition, 2005.

[95] A. Shokrollahi. Raptor Codes. *IEEE/ACM Trans. Netw.*, 14(SI):2551–2567, 2006.

[96] Arlei Silva, Wagner Meira Jr., and Mohammed J. Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *CoRR*, abs/1201.6568, 2012.

[97] Thiago H. Silva, Pedro O.S. Vaz de Melo, Jussara M. Almeida, and Antonio A.F. Loureiro. Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pages 000528–000534, July 2013.

[98] Thiago H. Silva, Pedro O.S. Vaz de Melo, Jussara M. Almeida, and Antonio A.F. Loureiro. On the use of participatory sensing to better understand city dynamics. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, pages 1347–1350, New York, NY, USA, 2013. ACM.

[99] Thiago Henrique Silva, Pedro Olmo Stancioli Vaz de Melo, Jussara Marques de Almeida, and Antonio Alfredo Ferreira Loureiro. Uncovering properties in participatory sensor networks. In *Proceedings of the 4th ACM International Workshop on Hot Topics in Planet-scale Measurement*, HotPlanet '12, pages 33–38, New York, NY, USA, 2012. ACM.

[100] ThiagoH. Silva, Pedro O. S. Vaz Melo, AlineCarneiro Viana, JussaraM. Almeida, Juliana Salles, and AntonioA.F. Loureiro. Traffic condition is more than colored lines on a map: Characterization of waze alerts. In Adam Jatowt, Ee-Peng Lim, Ying Ding, Asako Miura, Taro Tezuka, Gaël Dias, Katsumi Tanaka, Andrew Flanagin, and BingTian Dai, editors, *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*, pages 309–318. Springer International Publishing, 2013.

[101] Joel Sommers, Paul Barford, Nick Duffield, and Amos Ron. Accurate and Efficient SLA Compliance Monitoring. In *Proc. ACM SIGCOMM*, 2007.

[102] Stoyan Stefanov. YSlow 2.0. In *CSDN SD2C*, 2008.

[103] Malgorzata Steinder, Ian Whalley, and David M. Chess. Server virtualization in autonomic management of heterogeneous workloads. *Operating Systems Review*, 42(1):94–95, 2008.

[104] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, August 2010.

[105] Michael Toomim, Xianhang Zhang, James Fogarty, and James A. Landay. Access control by testing for shared knowledge. In Mary Czerwinski, Arnold M. Lund, and Desney S. Tan, editors, *CHI*, pages 193–196. ACM, 2008.

[106] D. Turner, K. Levchenko, A. Snoeren, and S. Savage. California Fault Lines: Understanding the Causes and Impact of Network Failures. In *Proc. ACM SIGCOMM*, 2010.

[107] Marisa Affonso Vasconcelos, Saulo Ricci, Jussara Almeida, Fabrício Benevenuto, and Virgílio Almeida. Tips, dones and todos: Uncovering user profiles in foursquare. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 653–662, New York, NY, USA, 2012. ACM.

[108] Márcio L. A. Vidal, Altigran Soares da Silva, Edleno Silva de Moura, and João M. B. Cavalcanti. Structure-driven crawler generation by example. In Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin, editors, *SIGIR*, pages 292–299. ACM, 2006.

[109] Alex Borges Vieira, Rafael Barra De Almeida, Jussara Marques De Almeida, and SéRgio Vale Aguiar Campos. Simplyrep: A simple and effective reputation system to fight pollution in p2p live streaming. *Comput. Netw.*, 57(4):1019–1036, March 2013.

[110] Michael Welzl. *Network Congestion Control: Managing Internet Traffic (Wiley Series on Communications Networking & Distributed Systems)*. John Wiley & Sons, 2005.

[111] T. White. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 2009.

[112] Faber Henrique Zacarias Xavier, Lucas Maia Silveira, Jussara Marques de Almeida, Carlos Henrique Silva Malab, Artur Ziviani, and Humberto Torres Marques-Neto. Understanding human mobility due to large-scale events. In *Proc. Third conference on the Analysis of Mobile Phone Datasets*, pages 45–47, 2013.

[113] Faber Henrique Zacarias Xavier, Lucas Maia Silveira, Jussara Marques de Almeida, Artur Ziviani, Carlos Henrique Silva Malab, and Humberto Torres Marques-Neto. Analyzing the workload

dynamics of a mobile phone network in large scale events. In *Proceedings of the First Workshop on Urban Networking*, UrbaNe '12, pages 37–42, New York, NY, USA, 2012. ACM.

[114] Faber Henrique Zacarias Xavier, Lucas Maia Silveira, Jussara Marques de Almeida, Artur Ziviani, Carlos Henrique Silva Malab, and Humberto Torres Marques-Neto. Análise da mobilidade humana em eventos de larga escala baseada em chamadas de telefones celulares. In *Anais do Seminário Integrado de Software e Hardware (SEMISH)*, 2013.

[115] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 177–186, New York, NY, USA, 2011. ACM.

[116] A.X. Zhang, A. Noulas, S. Scellato, and C. Mascolo. Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *Social Computing (SocialCom), 2013 International Conference on*, pages 69–74, Sept 2013.

[117] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph. Tapestry: A Fault-tolerant Wide-area Application Infrastructure. *SIGCOMM Comput. Commun. Rev.*, 32(1):81–81, 2002.

[118] Elena Zheleva and Lise Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 531–540, New York, NY, USA, 2009. ACM.

Tabela 2: Cronograma Previsto para Execução das Atividades Relacionadas ao Desafio 1.

Objetivo / Atividade	**Quadrimestre**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Objetivo 1: Caracterização dos padrões de comportamento de usuários do Dropbox	•	•	•									
Objetivo 1: Criação de um modelo de comportamento dos usuários do Dropbox				•	•	•	•					
Objetivo 1: Caracterização de padrões de mobilidade de usuários inferidos de aplicações da Web		•	•	•	•							
Objetivo 1: Modelagem de padrões de mobilidade						•	•	•	•			
Objetivo 2: Modelagem da rede de compartilhamento de conteúdo no Dropbox								•	•	•	•	
Objetivo 2: Caracterização das redes de colaboração científicas	•	•	•									
Objetivo 2: Desenvolvimento de métricas que capturam os interesses dos usuários para fins de personalização da recomendação de tags			•	•	•							
Objetivo 2: Validação das métricas de interesses em mecanismos de recomendação de tags						•	•	•	•	•		
Objetivo 3: Caracterização da importância relativa de vários atributos para a popularidade de um conteúdo	•	•										
Objetivo 3: Caracterização dos padrões de evolução de popularidade			•	•	•							
Objetivo 3: Desenvolvimento de modelos de evolução de popularidade de um conteúdo onlines						•	•	•	•			
Objetivo 3: Desenvolvimento de modelos de previsão de popularidade									•	•	•	•
Objetivo 4: Investigar o impacto da infiltração de robôs em redes sociais	•	•	•									
Objetivo 4: Investigar a manipulação da opinião dos usuários de sistemas sociais na Web				•	•	•	•					
Objetivo 4: Detecção e combate a usuários maliciosos								•	•	•	•	•
Objetivo 5: Geração de modelo de análise de projeto de configurações	•	•	•	•								
Objetivo 5: Geração de um modelo descritivo de elementos de sociabilidade				•	•	•	•	•				
Objetivo 5: Análise de métodos de avaliação de qualidades de uso em sistemas colaborativos									•	•	•	•
Objetivo 5: Análise interdisciplinar de privacidade em redes sociais				•	•	•	•	•	•			

Tabela 3: Cronograma Previsto para Execução das Atividades Relacionadas ao Desafio 2.

Objetivo / Atividade	**Quadrimestre**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Objetivo 1: Modelagem da estrutura base para representação de informação	•	•										
Objetivo 1: Desenvolvimento de mecanismos de indexação e recuperação		•	•	•								
Objetivo 2: Desenvolvimento de políticas de coleta de redes sociais				•	•	•						
Objetivo 2: Desenvolvimento de políticas de coleta de dados acadêmicos				•	•	•						
Objetivo 3: Desenvolvimento de mecanismos para integração de dados						•	•	•				
Objetivo 3: Desenvolvimento de modelos de qualidade de dados						•	•	•				
Objetivo 4: Desenvolvimento de mecanismos para enriquecimento textual						•	•	•				
Objetivo 4: Desenvolvimento de mecanismos para enriquecimento semântico						•	•	•				
Objetivo 5: Desenvolvimento de mecanismos para busca em redes semânticas								•	•	•		
Objetivo 5: Desenvolvimento de mecanismos para recomendação em dispositivos móveis										•	•	•
Objetivo 5: Desenvolvimento de mecanismos para descoberta de padrões em redes acadêmicas								•	•	•		
Objetivo 5: Desenvolvimento de mecanismos para descoberta de padrões em fluxos de dados										•	•	•

Tabela 4: Cronograma Previsto para Execução das Atividades Relacionadas ao Desafio 3.

Objetivo / Atividade	**Quadrimestre**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Objetivo 1: Extensão e desenvolvimento de técnicas de monitoramento	•	•	•	•	•	•						
Objetivo 1: Caracterização de dados de rede, centros de processamento e aplicações				•	•	•	•	•	•			
Objetivo 1: Modelagem do impacto de anomalias no desempenho de aplicações Web							•	•	•	•	•	•
Objetivo 2: Identificação e modelagem de infra-estruturas distribuídas	•	•	•									
Objetivo 2: Mapeamento de requisitos de serviços Web				•	•	•						
Objetivo 2: Desenvolvimento de arcabouço para prototipagem de serviços Web	•	•	•	•	•	•	•	•	•			
Objetivo 2: Implementação e avaliação de serviços Web descentralizados							•	•	•	•	•	•
Objetivo 3: Escalonamento em ambientes heterogêneos	•	•	•	•								
Objetivo 3: Reconfiguração dinâmica			•	•	•	•						
Objetivo 3: Programação em alto nível						•	•	•	•	•		
Objetivo 3: Suporte a profiling									•	•	•	•
Objetivo 4: Seleção de algoritmos a serem paralelizados	•	•	•	•								
Objetivo 4: Apropriação dos algoritmos			•	•	•	•	•					
Objetivo 4: Projeto dos algoritmos paralelos							•	•	•	•		
Objetivo 4: Implementação dos algoritmo paralelos									•	•	•	•
Objetivo 4: Validação dos algoritmos										•	•	•
Objetivo 5: Investigação de como e em que contextos algoritmos de análise de sentimentos podem apoiar usuários na seleção da informação desejada.	•	•	•	•	•	•						
Objetivo 5: Uso de simulação na antecipação dos resultados de configurações possíveis				•	•	•	•	•				
Objetivo 5: Análise da aplicabilidade de uma ontologia para recomendação de visualizações para dados qualitativos							•	•	•	•	•	•
Objetivo 5: Investigação de qualidade de uso e técnicas de interação								•	•	•	•	•

Laboratórios de Pesquisa
1. Lab. de Análise e Modelagem de Desempenho de Sistemas de Computação - CAMPS
2. Lab. de Banco de Dados - LBD
3. Lab. de Computação Paralela - LCP
4. Lab. para Tratamento da Informação - LATIN
5. Núcleo de Processamento Digital de Imagens - NPDI
6. Lab. de Vídeo sob Demanda- VoD

Tabela 5: Laboratórios de pesquisa e desenvolvimento que compõem o MASWeb.