Busca Contextualizada Enriquecida com Dados Abertos para Apoiar a Aprendizagem Colaborativa em Redes Sociais

Title: Contextual Search Enriched with Open Data for Supporting Collaborative Learning in Social Networks

Eduardo Fritzen

Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
eduardo.fritzen@uniriotec.br

Sean W. M. Siqueira

Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
sean@uniriotec.br

Leila C. V. Andrade

Universidade Federal do Estado do
Rio de Janeiro (UNIRIO)
leila@uniriotec.br

Resumo *Com o desenvolvimento das tecnologias de informação e comunicação, em especial a Web e sistemas colaborativos, observa-se um amplo acesso a motores de busca e redes sociais. Este trabalho considera que a relevância na recuperação de documentos na Web pode ser primordial para estimular a colaboração em dinâmicas de aprendizagem baseadas em discussão. Para melhorar e contextualizar a recuperação de documentos na Web, este artigo propõe: (i) a modelagem do contexto a partir da extração das mensagens em grupos de rede social e (ii) uso do contexto para melhorar a relevância na recuperação de documentos na Web. Um estudo de caso demonstrou que a captura do contexto enriquecido usando mensagens de discussão pode melhorar a relevância dos resultados das buscas na Web e contribuir com as discussões.*

Palavras-Chave: Busca contextual, Dados abertos, Aprendizagem Colaborativa, Redes Sociais

Abstract *With the development of information and communication technologies, in special the Web and collaborative systems, there is a broad access to query engines and social networks. This work considers the relevance in documents retrieval on the Web can be important to encourage collaborative learning dynamics based on discussion. To improve and contextualize the documents retrieval on the Web, this paper proposes: (i) the context modeling from the extraction of messages in social networking groups and (ii) use of the context to improve the relevance in the documents retrieval on the Web. A case study showed the enriched context capture using the discussion messages can improve the relevance of search results on the Web and contribute to the discussions.*

Keywords: Contextual search, Open data, Collaborative Learning, Social Networks

1 Introdução

Os novos recursos e os avanços das aplicações disponíveis na *Web* influenciam cada vez mais o cotidiano das pessoas [1]. As novas aplicações, chamadas de mídias sociais¹, são cunhadas sobre as tecnologias e ideologias da *Web Social*². Neste contexto, há a formação de redes sociais online, que merecem destaque pela sua popularidade ascendente tanto em número de usuários, quanto ao tempo que esses dedicam à navegação. Dos brasileiros que acessam a internet, 72% já incorporaram às suas rotinas o hábito de navegar em pelo menos um sítio de rede social online, dos quais 69,3% se enquadram na faixa etária entre 16 e 35 anos [4].

Sítios de redes sociais geralmente fornecem recursos para o compartilhamento de conteúdo, como a publicação de documentos e *links*, bem como a troca de mensagens usando programas de comunicação. Estas funcionalidades permitem o uso destes sítios como um ambiente oportuno à aprendizagem colaborativa, sendo que alguns permitem criar grupos de usuários que compartilham o mesmo interesse (por exemplo, alunos de uma classe ou um curso). Em sítios de redes sociais usados para o ensino-aprendizagem, o compartilhamento de conteúdo permite fornecer o material de aprendizagem necessário para um curso, enquanto o programa de comunicação permite a troca de ideias.

O fato de muitos estudantes já usarem sítios de redes sociais fez com que professores comesçassem a se familiarizar com esta tendência para usá-la a seu favor, explorando novas possibilidades para melhorar os resultados da aprendizagem. Existem muitos exemplos de sítios de redes sociais usados por professores e alunos como provedores de comunicação [5, 6, 7, 8].

Além das mídias sociais, é importante destacar que *Web* é uma gigantesca fonte de informação que cresce de forma acelerada, como demonstra estimativas da Netcraft [9], que indica que existem mais de 500 milhões de sítios no mundo. Deste modo, uma tecnologia amplamente difundida e utilizada na navegação dos usuários na *Web* corresponde aos motores de busca [10]. Motores de busca na *Web*, tais como Google³ e Bing⁴, foram concebidos para possibilitar a recuperação da informação desejada em meio às dezenas de milhões de páginas disponíveis na *Web*.

Entretanto, ao observar o resultado dos motores de buscas disponibilizados nos dois sítios de redes sociais mais populares do Brasil, Facebook⁵ e Orkut⁶ [11], percebe-se que o Orkut restringe os resultados da busca a elementos de sua própria rede, como pessoas e comunidades, enquanto no Facebook os resultados da *Web* são idênticos aos obtidos com o sítio de seu motor de busca *Web*, independente se a pesquisa originou-se no sítio do buscador ou a partir de um grupo de discussão no sítio da rede social. Este problema se torna mais crítico quando as redes sociais envolvem atividades de trabalho ou de ensino-aprendizagem, pois as pesquisas na *Web* realizadas a partir destes ambientes visam auxiliar o desenvolvimento dessas atividades. Melhorar a precisão dos resultados das buscas pode estimular as discussões no grupo e promover a colaboração. Por exemplo, em um ambiente educacional, a relevância dos documentos obtidos a partir da *Web* pode apoiar o aprendizado, uma vez que o conteúdo retornado será possivelmente mais adequado à necessidade de informação do aprendiz.

Entretanto, em geral, a expressão de busca (conjunto de palavras-chave) informada pelo usuário é composta por poucos termos. Segundo levantamento feito pela empresa Experian Hitwise, entre 26/08/2011 e 26/11/2011, 66,55% das buscas utilizam no máximo três palavras [12]. Além da quantidade, a qualidade dos termos também deve ser considerada, frente a dificuldade que grande parte dos usuários tem em definir quais palavras-chave representam os documentos de seu interesse e que deverão compor a sua expressão de busca [13]. Esta expressão de busca é, portanto, pouco representativa do contexto do domínio do usuário e suscetível a ambiguidades que degradam o processamento e resultado da busca (ou seja, os documentos que compõem o resultado não são tão relevantes para o usuário de acordo com seu contexto).

O uso de informações relacionadas ao contexto pode melhorar a relevância dos resultados a partir de ajustes na consulta do usuário. Estes ajustes podem ser, por exemplo, através do uso da técnica de expansão de consultas. A técnica de expansão de consultas, citada na área de recuperação de informação [14] e adotada neste trabalho, consiste em adicionar termos à consulta original, a fim de diminuir a ambiguidade e promover maior acurácia nos resultados. Quanto mais termos, e mais representativos estes termos forem, maior a possibilidade de encontrar documentos relevantes [15].

A intenção é tornar a recuperação de informação sensível ao contexto das discussões (por intermédio do uso das mensagens para modelagem do contexto), oferecendo, portanto, resultados de busca contextualizados. Con-

¹ Segundo Kaplan e Haenlein [2], Mídia Social é um grupo de aplicações baseadas na *Web* que permitem a criação e a troca de conteúdo digital gerado pelo usuário.

² Segundo Porter [3], *Web Social* (ou *Web 2.0*) é um conjunto de relações entre pessoas na *Web* com suporte de aplicações projetadas para apoiar e fomentar a interação social.

³ <http://www.google.com>

⁴ <http://www.bing.com>

⁵ <http://www.facebook.com>

⁶ <http://www.orkut.com.br>

tudo, sítios de redes sociais possuem características que dificultam a captura do contexto, como ausência de conteúdo no início das discussões e mensagens escritas de maneira informal, com o uso de abreviações e linguagem coloquial, expressas com poucas palavras. Para transpor esse obstáculo, é proposta a captura do contexto a partir do enriquecimento destas mensagens em dados abertos e o uso deste contexto para melhorar a consulta do usuário, fornecendo-lhe conteúdos mais relevantes a partir da *Web*.

Assim, este artigo apresenta uma arquitetura para extração do contexto de discussões em redes sociais e seu enriquecimento a partir de dados abertos. O enriquecimento objetivou resolver o problema de pouca informação nas discussões, principalmente no início das mesmas (poucos termos). Tem-se como objetivo geral a melhoria dos resultados das pesquisas com o uso da técnica de expansão de consultas a partir de discussões, com termos cujo significado foi enriquecido com base no conteúdo de uma enciclopédia *Wiki*.

O restante deste trabalho está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados, a seção 3 apresenta a proposta para captura de contexto durante discussões, a seção 4 detalha o estudo de caso, sua metodologia e forma de avaliação, a seção 5 as métricas utilizadas para avaliação, a seção 6 apresenta os resultados a partir das métricas e a seção 7 traz as considerações finais.

2 Trabalhos Relacionados

Este artigo relaciona-se com trabalhos que, independente de técnica, usam algum modelo de conhecimento para tornar a recuperação da informação sensível ao contexto. Informações de domínio do conhecimento, de processo de negócios, fornecidas pelo usuário (explicitamente) ou obtidas a partir de seus comportamentos (implicitamente) podem ser usadas para modelar o contexto. Esta modelagem pode exigir grande esforço humano, como a criação de ontologias por especialistas, preenchimento de preferências e marcação de documentos (modelagem manual) ou técnicas computacionais que supram (modelagem automática) ou minimizem (modelagem semiautomática) o esforço humano, como a análise de cliques, *corpus* textual, dados históricos, sensores ou outras maneiras ubíquas [16, 17].

A técnica de mineração de dados educacionais (*Educational Data Mining*¹ - EDM) é uma disciplina científica que visa o desenvolvimento de métodos para descoberta de conhecimento e padrões em grandes coleções de dados educacionais [18, 19]. Alguns trabalhos sobre técnicas de EDM podem ser encontradas em [20, 21]. No entanto,

esta abordagem baseia-se em grandes coleções oriundas de bases de dados de ambientes educacionais, que não é o caso deste trabalho que se baseia em poucas mensagens para a definição do contexto.

Rensing et al. [22] discutem a diferença entre os métodos baseados em conteúdo e sistemas de filtragem colaborativa para a recomendação de recursos de aprendizagem, propondo uma abordagem baseada em *tagging* colaborativo. Embora seja uma abordagem interessante, exige esforço dos participantes, que além de discutir e compartilhar conteúdo devem marcar os conteúdos nas discussões. Então, o conteúdo marcado nas discussões é processado, a fim de adquirir o contexto de aprendizagem.

Zhuahdar e Nasraoui [23] capturaram o contexto com o uso de taxonomias de domínio e perfis de usuário, e reclassificaram os resultados da pesquisa de acordo com a similaridade entre os termos contidos nos resultados da pesquisa e taxonomias.

Paula [24] propõe o uso de uma ontologia de domínio (ciência da computação) para apoiar a recuperação de informação, usando a técnica de expansão de consultas para recuperar documentos em uma coleção experimental composta por 889 artigos relacionados a subáreas de ciência da computação.

Ambrósio et al. [25] utilizaram técnicas de mineração de textos em um conjunto de documentos (apresentações) para recomendar documentos armazenados em um repositório e também sugerir a expansão da consulta a buscas na *Web*.

Kang et al. [26] realizaram o agrupamento dos *x* primeiros documentos retornados pela consulta original e extraíram os termos mais relevantes de cada agrupamento. O usuário deve selecionar um dos *x* agrupamentos sugeridos, de acordo com o que julgar relevante. Os termos do agrupamento selecionado são somados aos termos da consulta original e o resultado da consulta expandida é apresentado ao usuário.

Prates e Siqueira [27, 28] propõem um método semi-automático para apoiar a contextualização das atividades de busca na *Web*. Sua abordagem realiza a extração de conhecimento dependente de *corpus* (análise e processamento textual) para então utilizar técnica de expansão das consultas e as executa em motores de busca *Web*. A abordagem para a expansão de consulta assume que os termos mais frequentes em documentos que são representativos de um domínio têm maior probabilidade de ocorrer em sítios e documentos disponíveis na internet e são relevantes e relacionados a este domínio. Além disso, considera que um documento pode tratar de diversos assuntos e com isso, a criação da lista de termos, usada na expansão das consultas, faz uso de técnicas de segmenta-

¹ <http://www.educationaldatamining.org>

ção de tópicos por assuntos antes da etapa de agrupamento (*clustering*), aplicadas a um conjunto de arquivos de contexto representativos do domínio. O usuário pode visualizar e usar para a expansão os termos de todos os *clusters*, não só os termos do *cluster* que mais se relaciona aos termos da consulta.

Na literatura especializada é possível encontrar muitas propostas que visam a melhoria da recuperação da informação com o uso de expansão de consultas sensíveis ao contexto, como a proposta deste artigo. Foram descritos alguns desses trabalhos, de modo especial, aqueles que empregam o *corpus* textual para a geração do modelo de contexto e, a partir deste modelo, propõem a recuperação da informação mais adequada às necessidades do usuário. A abordagem apresentada neste artigo difere das demais na origem dos dados (discussões em plataformas de rede social) e na forma em que o contexto é modelado (enriquecimento das discussões em dados abertos).

3 Enriquecimento de termos de discussões em redes sociais

A arquitetura para a geração do contexto parte somente das discussões realizadas em uma rede social. É importante ressaltar que as extrações do contexto são realizadas no momento em que as discussões ocorrem (em tempo real), para que os resultados das buscas possam estar atualizados com as últimas postagens. Para a dinâmica, os alunos foram incitados ao debate e, embora não proibitivo, o envio de *links* foi desestimulado e desconsiderado para fins de extração de informação. Portanto, a proposta de arquitetura considerou: (i) o fato de não existir documentos (normalmente composto por muitas palavras), mas somente mensagens (normalmente compostas por poucas palavras), (ii) não haver pré-processamento do contexto e (iii) o ambiente de discussão estar em constante modificação.

O enfoque de solução usa mensagens de softwares de comunicação, como as discussões em grupos das redes sociais para gerar o contexto e usá-lo para extrair e sugerir palavras-chave que mais se aproximem da expressão de busca informada pelo usuário. Estas palavras são enriquecidas a partir de dados abertos e poderão ser combinadas pelo usuário e adicionadas à consulta original. Pode-se dizer então que a consulta é expandida a partir da construção de vocabulários gerados automaticamente por algoritmos de processamento de texto, com a ajuda do usuário.

O enriquecimento do contexto é importante para acrescentar termos correlatos, visto que, em geral, a quantidade de texto presente nas mensagens em redes sociais é pequena se comparada a outros conteúdos digitais, como livros, dissertações, artigos, apostilas, páginas

da *Web* etc. Portanto, o simples uso de medidas de frequência para extrair termos relevantes em uma discussão pode não ser suficiente para prover bons resultados, principalmente no início das discussões. Assim, este trabalho deve lidar com problema do arranque a frio (*cold start problem*). Tal como Chen e Sycara [29], Smrž e Schmidt [30], e Narayan [31], o caráter incremental para a tarefa de extração de informação é considerado, por operar sobre um ambiente que cresce gradualmente, como os grupos de discussão em plataformas de rede social.

Para o processamento das consultas, foram pesquisados algoritmos que analisassem a semelhança conceitual entre a expressão de busca e o contexto enriquecido. Isso faz com que os termos sugeridos sejam sensíveis ao contexto de domínio e relacionados à expressão de busca do usuário.

De modo a facilitar o entendimento da solução, a arquitetura foi organizada em três níveis de abstração: conceitual, lógico e físico.

3.1 Arquitetura Conceitual

A arquitetura conceitual (Figura 1) deve atender à premissa de que, a partir de informações do contexto pobre (como as mensagens de discussão) e de um conjunto de informações (ou documentos) que possa enriquecer o contexto (ou seja, uma base de conhecimento adicional), é possível expandir consultas *Web* do usuário a fim de obter recursos relacionados ao assunto discutido.

Deve-se observar que, muitas vezes, encontrar os recursos desejados é uma tarefa difícil, porque, em geral, os usuários não expressam com precisão sua necessidade de informação e os motores de busca não se adaptam a diferentes interesses de usuários ou grupos de usuários. Para se evitar esse problema, o contexto deve evoluir de acordo com as novas necessidades de informação do usuário, sob a pena de perda de eficiência na busca e recuperação de informação. Para tanto, a atividade de enriquecimento do contexto deve ocorrer continuamente a partir da modificação nas informações do contexto e serve como insumo para as buscas dos usuários.

Figura 1: Arquitetura Conceitual

3.2 Arquitetura Lógica

Para facilitar a explicação da arquitetura lógica, a mesma foi dividida em enriquecimento do contexto e processamento da consulta do usuário.

Enriquecimento do Contexto

O papel do processamento do contexto é gerar o contexto de domínio enriquecido a partir da seleção de informações que possam apoiar este enriquecimento, como artigos de uma base de conhecimentos relacionados às mensagens de discussão. Assim, neste trabalho, o contexto enriquecido é definido a partir da análise linguística e apoio semântico das discussões aliados a uma base conhecimento *Wiki*.

O contexto é gerado de acordo com as discussões, ou seja, somente mensagens são usadas para a construção do contexto. O conteúdo dos links publicados no grupo não foi considerado. Para tanto, as seguintes etapas devem ser contempladas pela arquitetura (Figura 2).

Figura 2: Arquitetura Lógica para Enriquecimento de Contexto

- Seleção de Mensagens – obtém as mensagens do grupo e as replica em uma base local denominada “Contexto Pobre”, também repassando ao Processamento Linguístico para a geração do “Contexto Enriquecido”.
- Processamento Linguístico – obtém termos nas discussões que se relacionem com artigos da *Wiki*.
- Enriquecimento – com o apoio de uma ontologia que descreve a enciclopédia *Wiki*, obtém o artigo (ou página *Wiki*) da etapa anterior bem como artigos relacionados.
- Processamento do Índice – organiza a informação obtida em um índice que garanta a rápida recuperação por semelhança entre o conteúdo do índice e uma expressão de busca.

Processamento da Consulta

O processamento da consulta usa as mensagens e o contexto enriquecido para sugerir os termos que poderão ser usados pelo usuário para expandir a sua expressão de busca por documentos na *Web*. Os termos sugeridos são aqueles que de alguma maneira se relacionam à discussão. Portanto, a seleção dos termos para expansão é influenciada pela comunicação entre os usuários, ou seja, pelos eventos de colaboração realizados no ambiente de discussão, modificando-se com o tempo. Com isso, os termos mais relevantes no tempo-1 talvez não sejam os mesmos no tempo-2, devido ao dinamismo do ambiente de discussões. A Figura 3 ilustra os componentes da

arquitetura lógica para o processamento da consulta do usuário.

Figura 3: Arquitetura Lógica para Processamento da Consulta

Primeiro, o usuário indica a necessidade de informação, indicando uma expressão de busca. O componente de seleção de consultas deve identificar estas solicitações de busca, para então proceder à extração das sugestões dos termos candidatos à expansão da consulta. A extração é feita a partir da análise da expressão de busca original e considera a similaridade semântica latente entre os termos da consulta e os documentos do contexto enriquecido. A Análise Semântica Latente é uma proposta para melhorar os problemas encontrados em abordagens de recuperação de informação baseada em palavras-chave, como o modelo de espaço vetorial (VSM – Vector Space Model). No modelo VSM, a matriz de documento possui, geralmente, dimensão elevada e esparsa, visto que cada palavra do *corpus* ocorre em poucos documentos deste mesmo *corpus*. Ainda, segundo Rodrigues e Asnani [32], a recuperação baseada em palavras-chave tem problemas em captar a estrutura semântica subjacente desejada e, com isso, torna-se vaga e suscetível a ruídos. Em outras palavras, estas consultas podem retornar documentos irrelevantes ou excluir do conjunto de respostas documentos potencialmente relevantes penalizados por não conter pelo menos uma palavra-chave da consulta.

Três grupos de termos são extraídos de maneiras distintas, de modo a possibilitar uma comparação da melhor estratégia para a captura do contexto e, portanto, da expansão de consulta.

O “Grupo A” é composto por rótulos (títulos) dos ar-

tigos de uma enciclopédia *Wiki* que apresentem em seu conteúdo maior grau de similaridade aos termos que compõem a expressão de busca. O “Grupo B” é composto por termos com maior ocorrência, a partir dos conteúdos dos artigos retornados no “Grupo A”. O objetivo deste grupo é possibilitar a sugestão de outros termos que sejam importantes, mas que não ocorram nos rótulos dos artigos. Obter o conteúdo dentro dos documentos visa ampliar a completude terminológica, não se restringindo a apenas ao seu rótulo. Por exemplo, o artigo cujo rótulo é “Banco de Dados” possui os termos {Atomicidade, Consistência, Isolamento e Durabilidade} em seu conteúdo. O componente “Seleção Contexto Enriquecido”, é responsável pela extração dos termos dos Grupos “A” e “B”. O componente “Seleção Contexto Pobre” é responsável pela extração dos termos do “Grupo C” e utiliza somente as mensagens de discussão para a extração dos termos e independe da consulta do usuário.

Já o componente “Geração Requisição” é responsável por agrupar os termos extraídos e gerar o acesso à busca. Uma requisição é o resultado de uma solicitação de consulta que serve como linha de base para o protótipo de busca e avaliação dos resultados.

Por fim, o componente de busca usa os dados consolidados na requisição para sugerir termos para a expansão de consultas. Adotou-se a modalidade de expansão de consulta interativa, visto que permite maior controle do usuário para a escolha dos termos que irão compor a expressão de busca [33]. Segundo Carpineto e Romano [14], a tarefa de compor a busca de maneira interativa por humanos pode evitar que os resultados da consulta se distanciem da intenção da busca (problema conhecido como *query drift*). Segundo Kelly et al. [34], a sugestão de termos em uma consulta interativa pode ser especialmente importante nos casos em que os usuários efetuam busca em domínios sobre os quais eles têm pouco conhecimento ou familiaridade.

3.3 Arquitetura Física

A arquitetura física baseou-se nos componentes descritos na arquitetura conceitual e lógica, bem como nas tecnologias da *Web Semântica* e *Web Social*. Partiu-se das mensagens trocadas em uma rede social online e do conhecimento de uma enciclopédia *Wiki* (*Web Social*), considerando tecnologias e conceitos da *Web Semântica*, como o acesso à DBpedia, uma ontologia responsável por definir conceitos e relações entre os mesmos.

A enciclopédia colaborativa *Wiki* considerada neste trabalho foi a *Wikipédia*, motivado pelo trabalho apresentado por Medelyan et al. [35], o qual observa algumas pesquisas que usam as informações da *Wikipédia* para auxiliar diversas áreas como o processamento de linguagem natural, recuperação de informação, extração de

informação e construção de ontologias.

A arquitetura física apresenta os componentes utilizados no desenvolvimento do protótipo em linguagem Java, baseado exclusivamente em componentes de uso gratuito e bibliotecas de código aberto. Foram pesquisadas as bibliotecas que melhor se enquadravam aos requisitos de desempenho relacionados ao tempo de resposta para as solicitações das consultas. O sistema desenvolvido deve ser eficiente (tempo), extensível (integração a outras redes sociais) e flexível (parametrizações).

Similarmente à arquitetura lógica, optou-se por dividir a arquitetura física em processamento do contexto e processamento das consultas.

Extração e Enriquecimento do Contexto

A proposta de arquitetura física para a geração do contexto a partir da extração de mensagens em grupos de discussão é apresentada na Figura 4. A partir das mensagens realizou-se o processamento de linguagem natural a fim de identificar as entidades relacionadas e complementares ao contexto das discussões. Dois macrocomponentes podem ser identificados na proposta: “Extração do Contexto” e “Enriquecimento do Contexto”.

Figura 4: Componentes da Arquitetura Física do Processamento do Contexto

O objetivo da “Extração do Contexto” é obter o conteúdo textual das mensagens ou, especificamente, as postagens e os comentários de grupos do Facebook, incluindo informações de autoria, data e hora. Para tanto, o componente “Seleção de Mensagens” adotou a biblioteca RestFB, que usa o Open Graph para a extração de infor-

mação desta rede social. Toda mensagem extraída é submetida à tokenização, remoção de *stopwords* e *stemming*, para então ser armazenada em um banco de dados relacional MySQL pelo componente de persistência, excluindo-se apenas as mensagens de consulta e oriundas do agente assistente. O componente “Tokenização, Remoção de *Stopwords* e *Stemming*” realiza as tarefas de, respectivamente, (i) quebra do conteúdo das mensagens em *tokens*, (ii) remoção de palavras com pouca importância e (iii) redução das palavras resultantes a sua forma raiz. O conteúdo destas mensagens origina o Contexto “Pobre”, usado para que sejam extraídos termos do “Grupo C”.

O componente “Seleção de Mensagens”, além de fornecer conteúdo para a geração do Contexto “Pobre”, também é responsável pela seleção de conteúdo textual para o enriquecimento do contexto, antes da etapa de “tokenização, remoção de *stopwords* e *stemming*” (ou seja, considerou-se a mensagem em seu estado natural). Para o enriquecimento do contexto, o componente de seleção de mensagens deve ainda selecionar apenas as mensagens que possuam duas ou mais palavras, desconsiderando-se os caracteres especiais, de pontuação e as *stopwords*, visto que, em etapas posteriores não se considerarão os unigramas, ou seja, não se desejou tratar as aceções das palavras para fins de desambiguação (*Word Sense Disambiguation*). Exemplos: “;-)”, “<3”, “da hora!”, “legal!”, “ok...” etc.

O objetivo do “Enriquecimento do Contexto” é ampliar o contexto terminológico, enriquecendo-o a partir de dados abertos de uma enciclopédia colaborativa. Para tanto, três etapas devem ser cumpridas: processamento linguístico das mensagens, enriquecimento e processamento do índice semântico.

O “Processamento Linguístico” possui dois subcomponentes: “Etiquetagem Gramatical” e “Extração de Bigramas e Trigramas”. Na “Etiquetagem Gramatical”, todas as mensagens são submetidas ao processamento de linguagem natural para definição da classe gramatical das palavras (*Part of speech / lexical class*). Para a geração da etiquetagem utilizou-se o serviço F-EXT-WS treinado com o *corpus* em português Mac-Morpho [36]. O F-EXT-WS foi escolhido pelos seguintes motivos: (i) Suporte ao idioma português; (ii) uso livre, mediante cadastro; (iii) acesso padronizado via protocolo WSDL, compatível com uma vasta gama de linguagens, dentre elas o Java; (iv) fácil acesso à documentação e a exemplos de uso; (v) suporte às funções de NLP *part-of-speech tagging*; (vi) uso corporativo e referenciado por uma série de artigos, teses e dissertações; (vii) boa acurácia; e (viii) boa performance [37]. Alternativas ao F-EXT-WS, tais como JtextPro, OpenNLP, LingPipe, LX-Suite e NLTK foram analisadas e rejeitadas por uma série de motivos, tais como não oferecer suporte ao idioma português, licenciamento, falta de documentação e exemplos de uso.

Após a etiquetagem gramatical é realizada a identificação de bigramas e trigramas nas mensagens através do componente “Extração de Bigramas e Trigramas”. Os bigramas e trigramas serviram para identificar instâncias de artigos na nuvem da *Linked Data*, em particular a *DBpedia*, usadas para a geração do contexto enriquecido. Adotou-se a abordagem linguística para a detecção dos bigramas e trigramas (análise individual, mensagem a mensagem, visto que as mensagens possuem poucas palavras). A abordagem estatística de co-ocorrência não foi considerada uma boa alternativa, pois, em geral exige um *corpus* maior [38]. Usou-se o conjunto dos bigramas e trigramas identificados como parâmetro para a formulação da expressão de busca por instâncias em dados abertos (extrair informações por significado, visto que bigramas e trigramas são menos propensos a ambiguidades).

O “Enriquecimento” possui três subcomponentes: “Seleção de Instâncias”, “Extração de Instâncias Relacionadas” e “Extração de Conteúdo das Instâncias”. A “Seleção de Instâncias” tem por objetivo buscar todos os indícios de bigramas e trigramas na *DBpedia* para obter o apoio semântico em um vocabulário controlado. A proposta de arquitetura defende o princípio de que bigramas e trigramas encontrados em mensagens de discussão são representativos do domínio discutido e, portanto, são candidatos em potencial a aparecer em rótulos de instâncias na *DBpedia*. Palavras simples foram desconsideradas da solução na etapa de expansão de dados abertos, pois as relações entre as palavras simples possuem alto nível de ambiguidade como conotação (uso da palavra em sentido diferente do original), homônimos (significado diferente, mesma grafia) ou polissemia (muitos significados para a mesma palavra). A problemática que envolve o tratamento automático (ou semi-automático) de ambiguidade [39] é um campo de estudo reconhecidamente complexo [40, 41]. Além disso, observa-se que os bigramas e trigramas representam a maior parte dos artigos da *Wikipédia*.

Com o apoio da biblioteca Jena¹, buscou-se o mapeamento entre bigramas e trigramas a instâncias da *DBpedia*. Para tanto, utilizou-se o rótulo *rdfs:label* e consultas na *DBpedia* a partir do SPARQL [42]. Os rótulos *rdfs:label* dos recursos da *DBpedia* são criados a partir dos títulos das páginas da *Wikipédia*. As instâncias selecionadas são todas aquelas em que houver casamento (*matching*) entre os bigramas e trigramas encontrados no texto e os dados abertos.

O componente de “Extração de Instâncias Relacionadas” deve agregar à base contextual enriquecida as instâncias relacionadas àquelas obtidas pelo componente “Seleção de Instâncias”. Para tanto, utilizou-se a propriedade Dublin Core Subject (*dcterms:subject*) e o apoio da biblioteca Jena para o tratamento das triplas RDF.

Todas as instâncias obtidas pela relação de afinidade direta, ou seja, artigos de uma mesma categoria (*dcterms:subject*) foram consideradas. Categoria² (por exemplo, *Category:Management_systems*) é um tipo especial de recurso da *DBpedia*, que é extraído a partir da classificação e agrupamento de artigos na *Wikipédia* que tratam sobre um mesmo assunto. Pode ser visto como um conceito abstrato (*Concept*) que agrupa um conjunto de recursos (artigos) relacionados [43]. Cada recurso do tipo *Category* na *DBpedia* associa-se a um *Concept* SKOS³ (*Simple Knowledge Organization System*) por intermédio da propriedade `<rdf:type rdf:resource="skos:Concept">`.

O componente “Extração de Conteúdo das Instâncias” é responsável por obter o conteúdo na *Wikipédia* de todas as instâncias, tanto aquelas obtidas pela “Seleção de Instâncias”, quanto as obtidas pela “Extração de Instâncias Relacionadas”. Como a *DBpedia* não traz o conteúdo do artigo, a base da *Wikipédia* foi utilizada (na implementação optou-se por uma versão baixada (*dump*) de modo a evitar acessos constantes à *Wikipédia*, mas tal versão só precisou ser baixada uma vez e poderia ser atualizada de tempos em tempos). A base da *Wikipédia* foi, então, indexada em um repositório local com o apoio do Apache Lucene⁴, uma biblioteca de indexação e busca textual escrita em *Java*.

Por fim, é realizado o “Processamento do Índice Semântico”, composto por dois subcomponentes, “Tokenização, Remoção *Stopwords* e *Stemming*” e “Atualização do Índice Semântico”. Nesta camada, a “Tokenização, Remoção de *Stopwords* e *Stemming*” modifica o conteúdo de instâncias dos artigos obtidos pela “Extração de Conteúdo”. Os bigramas e trigramas são unidos e tratados como um único *token* no caso em que houver casamento entre estes termos compostos em dados abertos. Por exemplo, se o trigramma “rio de janeiro” ocorre em dados abertos, deve-se considerar o *token* “rio_de_janeiro” e, em seguida, ignorar a remoção da *stopword* (preposição “de”) e stemizar as demais palavras, resultando no *token* “rio_de_jan”. O ORENGO foi utilizado para o *stemming*, sendo realizado a partir do PTStemmer⁵, uma biblioteca de stemização para o idioma português [44].

Finalmente, a subtarefa de “Atualização do Índice Semântico” deve ocorrer para que o contexto enriquecido possa ser pesquisável por intermédio da técnica de análise semântica latente. Para tanto, usou-se a biblioteca *Semantic Vectors*⁶ para explorar o espaço semântico textual e resolver problemas de relacionar a expressão de busca a um conjunto de documentos que compõem o contexto enriquecido. O *Semantic Vectors* usa os dados indexados

¹ <http://incubator.apache.org/jena/>

² <http://en.wikipedia.org/wiki/Help:Category>

³ <http://www.w3.org/2004/02/skos/>

⁴ <http://lucene.apache.org/java/docs/>

⁵ <http://code.google.com/p/ptstemmer/>

⁶ <http://code.google.com/p/semanticvectors/>

no Lucene para a geração da matriz de termos e documentos.

Processamento da Consulta e Extração dos Termos

A Figura 5 ilustra os componentes utilizados na arquitetura física do “Processamento da Consulta”, responsável pela expansão da consulta e geração da requisição de busca que será utilizada pelo protótipo busca na *Web*. A expressão de busca é capturada pelo componente “Seleção de Consulta”, que utiliza a biblioteca *RestFB* para comunicação com o *Facebook*. Apenas as mensagens marcadas como expressões de busca são selecionadas. Os termos informados na expressão de busca são submetidos ao processo de “Tokenização, Remoção de *Stopwords* e *Stemming*”.

Figura 5: Componentes da Arquitetura Física da Busca

O componente “Seleção do Contexto Enriquecido” utiliza a biblioteca *Semantic Vectors* para associar os termos presentes na consulta do usuário a um conjunto de documentos do contexto enriquecido e seus respectivos rótulos. Por questões de limitação da quantidade de termos a serem exibidos para os usuários, para esta pesquisa foram considerados apenas os seis (6) documentos com maior *score* entre termos da consulta e documentos do índice, obtido pelo algoritmo de LSA do pacote *Semantic Vectors*. Utilizou-se o rótulo dos documentos para a geração das sugestões do “Grupo A”. Estes rótulos foram transformados, de acordo com uma regra de formação. Os rótulos com até três (3) palavras, devem aparecer entre aspas (busca com a frase exata). Utilizou-se a busca por frase exata até três termos, pois este modo de busca apresentou bons resultados em [45]. O mesmo é válido para as palavras dispostas nos parênteses de desambiguação

(quando existir). Por exemplo, o artigo “C (linguagem de programação)” usa a expressão entre parênteses (desambiguação) para diferenciar este artigo dos demais, como o artigo que trata da letra “C” do alfabeto. Neste exemplo, os termos dispostos entre parênteses são adicionados à expressão de busca entre aspas, para que a busca ocorra pela frase exata. Caso o rótulo seja composto por quatro (4) ou mais palavras, todas as palavras serão consideradas na expressão de busca de maneira individualizada, ou seja, sem restrição de ordem para os termos.

Para o “Grupo B” considerou-se as palavras mais frequentes contidas nos documentos retornados pelo “Grupo A”. A frequência dos termos foi obtida com a aplicação do logaritmo na fórmula de frequência de termos [46]. Cada artigo pode ser visto como um documento. O cálculo da frequência (*term frequency*) aliado à aplicação do logaritmo visa reduzir a influência negativa de possíveis artigos compostos por muitos termos (normalização). Caso o termo mais frequente fosse um bigrama ou trigramma, o mesmo seria tratado entre aspas.

Já o componente “Seleção do Contexto Pobre” obtém um instantâneo dos termos mais relevantes nas mensagens no momento da busca, ou seja, foram utilizados apenas os termos mais frequentes das discussões para a geração do “Grupo C”. A frequência dos termos foi obtida com a aplicação do logaritmo na fórmula de frequência de termos [46]. Cada mensagem pode ser vista como um documento.

O objetivo do componente “Geração de Requisição” é unificar todos os termos sugeridos nos grupos A, B e C, gravá-los em banco de dados e, em seguida, informar ao usuário solicitante que a busca está disponível.

Por fim, o componente “Busca” é usado pelo usuário para acessar os documentos *Web* e realizar as avaliações de relevância. Este componente representa a parte gráfica do protótipo, denominada CCS Agent (*Collaborative Context Search Agent*). Utilizou-se o padrão MVC (*Model View Controller*), composto por JSP (*Java Server Pages*) e HTML na interface, controlador Servlet que trata a entrada de dados e o modelo do domínio descrito por Entidades JPA (*Java Persistence Architecture*). A recuperação de informações na *Web* usou a *API Google Search* configurada para páginas em português do Brasil, seguindo a política de uso fornecida pela Google Inc. A biblioteca *Google API para Java*¹ fornece acesso a diversos serviços do Google, como o *AdSense*, *Calendar*, *Analytics*, *Freebase*, *Groups*, *Google+*, *Orkut*, *Translate* etc. A *CustomSearch API*² foi utilizada para obter resultados de busca personalizada do Google. A administração é feita pelo *Google Console API*³.

¹ <http://code.google.com/p/google-api-java-client/>

² <https://developers.google.com/custom-search/>

³ <https://code.google.com/apis/console>

3.4 O Protótipo CCS Agent

O protótipo de aplicação (CCS Agent) desenvolvido possibilita que a busca por documentos *Web* seja feita, a qualquer momento, com o auxílio de um agente assistente solicitado a partir da própria interface de discussão. A ideia do protótipo foi possibilitar uma resposta para uma busca na *Web* paralelamente às discussões na rede social, de maneira a auxiliar a construção do conhecimento desejado sobre o tema no momento que a discussão ocorre. Para isto, a necessidade de informação (que indicará os termos originais de busca) deve ser indicada na forma de uma expressão disposta entre sinais duplos de interrogação, conforme exemplificado no fluxo “C” da Figura 6.

O sistema deve ser capaz de processar a solicitação do usuário e devolver uma mensagem com o *link* para o protótipo, que sugere termos a serem incluídos na consulta e usa o Google como provedor de documentos *Web*. Assim o aluno evita acessar outro ambiente para realizar a consulta, enquanto aguarda a mensagem de resposta no próprio grupo. A intenção é manter o usuário no grupo enquanto o sistema processa sua requisição.

A Figura 6 ilustra a visão geral da proposta de solução para facilitar o entendimento do trabalho apresentado neste artigo. Duas macrofuncionalidades podem ser identificadas na figura: processamento do contexto e processamento da consulta do usuário. O processamento do contexto é alimentado por mensagens do grupo (setas A e B), ou seja, a colaboração entre os usuários irá gerar

conteúdo que servirá de insumo para a geração do contexto. Já o processamento da consulta inicia-se pela identificação de uma solicitação de consulta (C), processamento da requisição, que extrai os termos sugeridos para expansão e formula uma mensagem de resposta com o *link* para o protótipo associado à requisição gerada (D). O usuário deve clicar nesse *link* (E) para acessar a interface do protótipo de busca (que é personalizado para sua necessidade de informação), escolher os termos para expandir a consulta e avaliar os resultados da busca, clicando sobre as estrelas (F).

Cada nova requisição gera três conjuntos de termos (lado esquerdo do site de busca e avaliação, na interface do CCS Agent), extraídos de maneira distinta e vinculados à expressão de busca informada pelo usuário. A busca é realizada pela combinação dos termos selecionados e gera um conjunto de documentos como resultado. O usuário deve marcar todos os termos que julgar necessário para serem utilizados na expansão de consulta. As repetições de termos (devido ao fato de serem provenientes dos três grupos diferentes, “A”, “B” e “C”) são combinadas de modo a exibir o termo apenas uma vez.

Duas consultas são enviadas ao *Google Search API*. A primeira contém os termos originais da consulta e a segunda, além dos termos da consulta original, os termos escolhidos para expansão da consulta, deste modo tem-se o resultado de controle e o resultado da expansão para comparação.

Figura 6 : Visão Geral da Proposta

4 Estudo de Caso

Um estudo de caso único foi conduzido em um ambiente acadêmico (UNIRIO) e teve a participação de alunos na utilização de um protótipo. Participaram do estudo 18 alunos do primeiro período do curso de Bacharelado em Sistemas de Informação, turma BSI 2011.2, matriculados na disciplina “Fundamentos de Sistemas de Informação”. Os participantes possuíam idade entre 18 e 26 anos, média e mediana de 20 anos. Todos os participantes possuem computador e acesso a internet. Apenas um participante não possuía perfil na rede social Facebook, que foi prontamente criado para participar da pesquisa.

Os alunos foram instruídos a participar de uma aula baseada em discussão, realizada no Facebook. No protótipo, cada aluno pôde executar quantas buscas e combinações de termos que julgasse necessárias e avaliaram a relevância de todos os resultados retornados. Após a realização da dinâmica, os participantes foram convidados a preencher um questionário com questões qualitativas sobre a relevância dos resultados na utilização do protótipo CCS Agent. Do total de (18) participantes, 14 responderam o questionário, cujos resultados foram analisados qualitativamente e comparados aos resultados quantitativos extraídos do protótipo.

O estudo de caso foi realizado para observar e relacionar a relevância dos resultados retornados pelas duas modalidades de consulta (original e expandida) e a combinação de termos selecionada nos grupos A e B (termos do contexto com enriquecimento) ou C (termos do contexto sem enriquecimento). A relevância de cada documento retornado no protótipo é aferida explicitamente pelos usuários numa escala de cinco (5) estrelas e relaciona-se implicitamente com o(s) grupo(s) do(s) termo(s) selecionado(s). Em outras palavras, foi possível avaliar a relevância dos resultados da consulta original com os resultados da consulta expandida, observando-se o grupo de termos escolhido para a expansão da consulta (ou suas combinações). Foi obrigatória a escolha de pelo menos um termo para expansão, independente do grupo.

5 Métricas

As avaliações coletadas no protótipo representam julgamento humano de relevância para cada documento *Web* retornado. O resultado da consulta é composto por seis (6) documentos retornados a partir da consulta original e seis (6) documentos retornado a partir da consulta expandida. Para fins de consistência, os resultados são mesclados em uma única lista e os resultados redundantes são exibidos apenas uma vez e recebem a mesma avaliação. Caso os seis primeiros resultados de cada expansão (expressões de busca original e expandida) sejam diferentes, serão apresentados ao usuário doze resultados para

avaliação (valor máximo). Caso os primeiros seis primeiros resultados coincidam (caso extremo) para ambas as consultas (original e expandida), apenas seis resultados serão exibidos aos usuários (valor mínimo). As métricas, utilizadas para comparar os resultados da busca original e expandida, foram: (i) precisão total dos x primeiros resultados [47], (ii) comprimento da busca [48] e (iii) correlação de *ranking* [49].

A precisão total dos x primeiros resultados (*first x full precision*) mede a quantidade total de informações relevantes nos x primeiros documentos. O valor considerado para x foi seis (6), ou seja, considerou-se a precisão total dos seis (6) primeiros resultados.

O comprimento da busca (*search length*) mede o número de documentos não relevantes que um usuário deve examinar antes de encontrar uma quantidade x de documentos relevantes. Utilizou-se o parâmetro $x=2$ (os dois primeiros documentos consecutivos e relevantes) e considerou-se relevância igual a quatro ou cinco estrelas.

Por fim, correlação de *ranking* (*rank correlation*) mede a correlação entre a classificação do sistema de busca e o julgamento do usuário (escala de cinco (5) estrelas) para os resultados da busca. Como não existe acesso às notas reais de classificação atribuídas pelo sistema (Google no caso deste trabalho), utilizou-se a posição do documento no conjunto-resposta para presumir sua pontuação. Quanto maior a proximidade do documento ao topo da lista, maior a sua pontuação. Empregou-se o coeficiente de correlação de *Pearson* para o cálculo da correlação entre a matriz A , que representa as avaliações dos usuários e a matriz B , que representa a ponderação associada a sua posição no conjunto de resultados da busca. A matriz de correlação varia de 4 para o primeiro documento retornado (pontuação máxima atribuída pelo usuário a um documento) a 0 para o sexto documento retornado (pontuação mínima atribuída pelo usuário a um documento), com passo de 0.8. O resultado da correlação entre as variáveis é apresentado no intervalo $[-1, +1]$, onde os resultados próximos de $+1$ representam uma correlação perfeita positiva, 0 ausência de correlação linear e -1 correlação perfeita negativa entre as variáveis [50].

6 Resultados

Foram selecionadas 58 sugestões de termos (nos três grupos) para as 34 avaliações consideradas nos quatro (4) dias de estudo, uma média de 1,7 termos escolhidos por consulta. Em relação à métrica precisão total, a consulta expandida trouxe melhores resultados que a consulta original em todos os dias, conforme observado na Figura 7. As consultas expandidas apresentaram melhores resultados em 67% dos casos nos dias 2 e 5, 57% no dia 3 e 55% no dia 4. As consultas originais foram melhores que as consultas expandidas em 17% (dia 2), 36% (dia 3),

27% (dia 4) e 33% (dia 5). Com isso, a precisão média de pesquisa foi de 29% para a consulta original e de 62% para a consulta expandida. Assim, se observa que os resultados das consultas expandidas foram melhores na opinião dos alunos do que os provenientes das consultas originais.

Figura 7: Métrica Precisão Total por Dia de Pesquisa

Na segunda métrica, comprimento da busca, verificou-se a superioridade das consultas expandidas em relação às consultas originais. Houve superioridade em 41% dos resultados obtidos com as consultas expandidas e 18% dos resultados obtidos com a consulta original.

A última métrica, correlação de *ranking*, apresentou melhores resultados para a consulta original (62%) em relação à consulta expandida (29%), ou seja, a correlação entre as notas dos usuários e a ordem dos documentos retornados pelo motor de busca foi melhor para as consultas originais. Isto talvez possa ser justificado pela forma como o Google organiza os resultados para a exibição.

Observa-se que os somatórios para as métricas não totalizam 100% devido ao fato de empates não terem sido contabilizados.

Ao desmembrar todas as possibilidades de combinação entre os grupos A, B e C, surgem sete (7) possibilidades. A combinação e a quantidade de vezes que a mesma ocorreu no estudo são descritas na Tabela 1.

Combinações	Quantidade
Somente A	18
Somente B	2
Somente C	4
A e B	3
A e C	2
B e C	3
A, B e C	2

Tabela 1: Combinação dos grupos de termos

Melhores resultados das consultas expandidas, obtidos pela métrica de precisão total em relação à consulta

original, foram apresentados nas consultas que usaram termos extraídos do contexto enriquecido (“Grupo A”, “Grupo B” e “Grupo A e B”). Apesar de a amostragem ser pequena para finalidade de resultados conclusivos, a expansão com termos do “Grupo A” (rótulos da DBpedia) apresentou resultados promissores frente aos demais, uma vez que tais termos foram os mais selecionados pelos alunos como representativos do contexto e portanto para a expansão das consultas.

7 Considerações Finais

O objetivo desta pesquisa foi melhorar a relevância dos resultados das buscas à *Web* a partir do tratamento de informação contextual obtida em mensagens de grupos de redes sociais online, em especial no contexto educacional, onde alunos interagem na rede social online e podem ter suas necessidades de informações satisfeitas através do próprio ambiente de rede social, promovendo melhores interações. Verificou-se a viabilidade da proposta de recuperação de informação contextual em um grupo de uma plataforma de rede social mundialmente conhecida (Facebook) em um contexto de aprendizagem baseada em discussão. Usou-se um conjunto de bibliotecas *open source* para interagir com Facebook e explorar com sucesso os recursos do grupo e o Google para a execução das buscas a conteúdo na *Web*.

A abordagem utilizada para a construção coletiva do conhecimento considerou somente discussões e troca de ideias com os colegas. Observou-se grande interação entre os participantes da dinâmica, que aproveitaram para discutir assuntos relacionados ao tema proposto. A correlação de *Pearson* para a matriz [número de consultas; número de comentários] obteve valor de correlação 0,79, ou seja, existe uma correlação positiva entre o número de consultas e o número de comentários por assunto, indicando que quanto mais os alunos fazem buscas na *Web*, maior é a interação resultante entre eles.

Por fim, este trabalho defendeu a criação do contexto de domínio de maneira semi-automática, ou seja, com o uso de recursos e tecnologias existentes, que independam de esforço humano adicional para a modelagem deste domínio. A modelagem de novas ontologias foi descartada uma vez que a engenharia de ontologias tem custo elevado e exige dedicação de especialistas e tempo para a modelagem. Criar modelos de representação e realizar marcações (anotações semânticas) a partir desses modelos é uma tarefa trabalhosa que pode requerer dedicação de especialistas, embora possa apresentar bons resultados se bem realizada.

Dentre os possíveis trabalhos futuros, destacam-se:

- Testar outras formas de organização e visualização dos termos de modo a facilitar o entendimento e

- consequentemente a seleção dos termos pelos usuários, por exemplo agrupando-os de forma hierárquica;
- Personalizar a geração do contexto por aluno, selecionando apenas as trilhas de mensagens que este tenha alguma participação (interesse);
 - Considerar pesos diferenciados para a geração dos termos, de acordo com (i) a trilha de mensagens em que a busca foi realizada e (ii) priorização de laços sociais entre o solicitante da consulta e sua rede;
 - Avaliar a interação de cada aluno com o sistema e com os demais alunos e investigar se houve melhoria na aprendizagem (questão em aberto);
 - Avaliar os termos sugeridos a cada consulta, possibilitando pesos diferenciados aos termos de acordo com as avaliações recebidas em consultas passadas;
 - Utilizar outros algoritmos para relacionar os termos da consulta com os documentos que compõem o contexto enriquecido, como a projeção aleatória [51] e a indexação aleatória reflexiva [52];
 - Realizar um novo estudo de caso com uma amostragem maior e em outros domínios que não informática, tais como História, Música, Religião etc.;
 - Habilitar o suporte a outros ambientes, como o Google+ ou Moodle por exemplo.

Agradecimentos

Os autores agradecem aos alunos que participaram dos estudos de caso, a FAPERJ (E-26/101.509/2010 BOLSA/BBP e E-26/170028/2008 Programa INC&T) e CNPQ (557.128/2009-9, INCT on Web Science).

Referências

- [1] The Boston Consulting Group (BCG). Clicks Grow Like BRICS: G-20, Internet Economy to Expand at 10 Percent a Year Through 2016. <http://www.bcg.com/media/PressReleaseDetails.aspx?id=tcn:12-100468>, Jul. 2013
- [2] A. M. Kaplan, M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59-68, 2010. doi:10.1016/j.bushor.2009.09.003
- [3] J. Porter. Designing for the Social Web. New Riders, Indianapolis, 2008.
- [4] IBOPE Mídia. Brasileiros caem na rede social. <http://www.ibope.com.br/pt-br/noticias/Paginas/Brasileiros%20caem%20na%20rede%20social.aspx>, Jul. 2013
- [5] A. Mora-Soto. Collaborative Learning Experiences Using Social Networks. In: *Proceedings of International Conference on Education and New Learning Technologies - EDULEARN09*, Barcelona. páginas 4260-4270, 2009.
- [6] A. F. U. Mansur, R. A. D. Carvalho, M. C. V. Biazus. Rede de Saberes Coletivos (RESA): Um Ambiente Complexo para Aprendizagem Acadêmica por Meio de Redes Sociais. In *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE)*, Aracaju, páginas 1284-1293, 2011.
- [7] S. Dotta. Uso de uma Mídia Social como Ambiente Virtual de Aprendizagem. In *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE)*, Aracaju, páginas 610-619, 2011.
- [8] Q. Wang, H. L. Woo, C. L. Quek, Y. Yang, M. Liu. Using the Facebook group as a learning management system: An exploratory study. *British Journal of Educational Technology*, 43(3):428-438, 2011. doi:10.1111/j.1467-8535.2011.01195.x
- [9] NETCRAFT. Web Server Survey. <http://news.netcraft.com/archives/2011/10/06/october-2011-web-server-survey.html>, Jul. 2013
- [10] ALEXA. The top 500 sites on the web. <http://www.alexa.com/topsites>, Jul. 2013
- [11] COMSCORE. It's a Social World: Top 10 Need-to-Knows About Social Networking and Where It's Headed. http://www.brandchannel.com/images/papers/534_comscore_wp_social_media_report_1212.pdf, Jul. 2013.
- [12] EXPERIAN HITWISE. Buscas com uma palavra são maioria. http://www.serasaexperian.com.br/release/noticias/2011/noticia_00728.htm, Jul. 2013
- [13] E. Ferneda. Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. Tese de doutorado, Universidade de São Paulo (USP), Dez 2003.
- [14] C. Carpineto, G. Romano. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Computing Surveys*, 44(1), artigo 1:1-50,

- 2012.
- [15] R. B. Yates, B. R. Neto. Modern Information Retrieval. 1 ed, Addison Wesley, New York, 1999.
 - [16] J. Bhogal, A. Macfarlane, P. Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866-886, 2007.
 - [17] V. Chanana, A. Ginige, S. Murugesan. Improving information retrieval effectiveness by assigning context to documents. In *Proceedings of the International Symposium on Information and Communication Technologies (ISICT 2004)*, Las Vegas, páginas 2239-2244, 2004.
 - [18] C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.D. Baker (Eds.) Handbook of Educational Data Mining. CRC Press, Boca Raton, 2012.
 - [19] O. Scheuer, B. M. McLaren. Educational Data Mining. In: N. M. Seel (Ed.). *Encyclopedia of the Sciences of Learning*. Springer, New York, páginas 1075-1079, 2012.
 - [20] R. Baker, K. Yacef. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining (JEDM)*, 1(1):3-17, 2009.
 - [21] C. Romero, S. Ventura. Preface to the Special Issue on Data Mining for Personalised Educational Systems. *User Modeling and User-Adapted Interaction*. 21(1-2):1-3, 2011.
 - [22] C. Rensing, P. Scholl, D. Böhnstedt, R. Steinmetz. Recommending and finding multimedia resources in knowledge acquisition based on Web resources. *Proceedings of the 19th International Conference on Computer Communications and Networks*, Zurich, páginas 1-6, 2008.
 - [23] L. Zhuhadar, O. Nasraoui. Semantic Information Retrieval for Personalized E-Learning. In *Proceedings of the IEEE Conf. on Tools with Artificial Intelligence (ICTAI 2008)*. Dayton, páginas 364-368, 2008.
 - [24] A. A. B. Paula. Uma Proposta para Expansão Semântica de Consultas Baseada em Ontologia de Domínio Específico. Dissertação de Mestrado, Centro de Estudos e Sistemas Avançados do Recife (C.E.S.A.R.), Mar 2010.
 - [25] A. P. Ambrósio, L.O. Silva, V.G. Neto. Automatic Retrieval of Complementary Learning Material for Slide Presentations. In *Proceedings of the International Conference on Interactive Computer Aided Blended Learning (ICBL 2009)*, Florianópolis, 2009.
 - [26] J. W. Kang, H. Kang, M. Ko, H. S. Jeon, J. Nam. A Term Cluster Query Expansion Model Based on Classification Information in Natural Language Information Retrieval. In *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, Sanya, páginas 172-176, 2010.
 - [27] J. C. Prates, S. W. M. Siqueira. Contextual Query based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Knowledge Management. In *Proceedings of the Americas Conference on Information Systems (AMCIS)*, AIS Electronic Library, Detroit, paper 198, 2011.
 - [28] J. C. Prates, S. W. M. Siqueira. Using educational resources to improve the efficiency of Web searches for additional learning material. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT)*, Athens, páginas 563-567, 2011.
 - [29] L. Chen, K. Sycara. WebMate: A Personal Agent for Browsing and Searching. In *Proceedings of the 2nd International conference on Autonomous agents (AGENTS'98)*, Minneapolis, páginas 132-139, 1998.
 - [30] P. Smrz, M. Schmidt. Information Extraction in Semantic Wikis. In *Proceedings of the 4th Workshop on Semantic Wikis*. Herssonissos, 2009, <http://ceur-ws.org/Vol-464/paper-08.pdf>, Jul. 2013
 - [31] N. Narayan. Advanced Intranet Search Engine. Dissertação de Mestrado, Mälardalen University, Ago 2009.
 - [32] R. Rodrigues, K. Asnani. Concept Based Search Using LSI and Automatic Keyphrase Extraction. In *Proceedings of the 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET '10)*, Goa, páginas 573-577, 2010. doi: 10.1109/ICETET.2010.100
 - [33] G. Kanaan, R. Al-Shalabi, S. Ghwanmeh, B. Bani-Ismael. Interactive and automatic query expansion: A comparative study with an application on Arabic. *Amer. J. Appl. Sciences*, 5(11): 1433-1436, 2008.
 - [34] D. Kelly, K. Gyllstrom, E. W. Bailey. A Comparison of Query and Term Suggestion Features for Interactive Searching. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Re-*

- trieval, Boston, páginas 371-378, 2009.
- [35] O. Medelyan, D. Milne, C. Legg, I. H. Witten. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9): 716-754, 2009.
 - [36] S. Aluisio, J. Pelizzoni, A. Marchi, L. Oliveira, R. Manenti, V. Marquiasavel. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In *Proceedings of the Computational Processing of the Portuguese Language (Propor 2003)*, Faro, páginas 110-117, 2003.
 - [37] E. L. R. Fernandes, R. L. Milidiú, C. N. Santos. Portuguese language processing service. In *Proceedings of the 18th International Conference on the World Wide Web (WWW 2009)*, Madrid, 2009.
http://wwwconference.org/www2009/pdf/submissions/wwwiberoamerica09_submission_1.pdf, Jul. 2013.
 - [38] H. J. Peat, P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5): 378-383, 1991.
 - [39] D. McCarthy, R. Koeling, J. Weeds, J. Carroll. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553-590, 2007.
 - [40] A. F. Smeaton. Information Retrieval: Still Butting Heads with Natural Language Processing. In: M. T. Pazienza (Ed.). *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Springer, New York, páginas 115-138, 1997.
 - [41] R. Navigli. Word sense disambiguation. *ACM Computing Surveys*, 41(2):1-69, 2009.
 - [42] P. Wang, J. Hu, H.-Jun Zeng, L. Chen, Z. Chen. Improving Text Classification by Using Encyclopedia Knowledge, In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, páginas 332-341, 2007.
 - [43] R. Mirizzi, A. Ragone, T. D. Noia, E. D. Sciascio. Ranking the Linked Data: The Case of DBpedia. In *Proceedings of the 10th International Conference on Web Engineering (ICWE 2010)*, Vienna, páginas 337-354, 2010.
 - [44] V. Orenço, C. Huyck. A stemming algorithm for the Portuguese language. In *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE 2001)*, Laguna de San Rafael, páginas 186-193, 2001.
 - [45] D. Johnson, V. Malhotra, P. Vamplew. More Effective Web Search Using Bigrams and Trigrams. *Webology*, 3(4):1-12, 2006.
 - [46] C.D. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, Cambridge, 2008.
 - [47] M. H. Chignell, J. Gwizdka, R. C. Bodner. Discriminating meta-search: A framework for evaluation, *Information Processing and Management: an International Journal*, 35(3):337-362, 1999.
 - [48] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Journal of American Society of Information Science*, 19(1):30-41, 1968.
 - [49] L. T. Su, H. L. Chen, X. Y. Dong. Evaluation of Web-based search engines from an end-user's perspective: A pilot study. In *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, Pittsburgh, páginas 348-361, 1998.
 - [50] M.G. Kendall, A. Stuart. The Advanced Theory of Statistics. Volume 2: Inference and Relationship. Hodder Arnold, Oxford. 1973.
 - [51] M. Sahlgren. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen 2005.
 - [52] T. Cohen, R. Schvaneveldt, D. Widdows. Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections, *Journal of Biomedical Informatics*, 43(2):240-256, 2010. doi: 10.1016/j.jbi.2009.09.003