Recuperação de Informações Textuais e Multimídia Utilizando Expansão de Consulta a Recursos da WEB 2.0

Vinícius Roggia Gomes[1], Paulo Henrique Guimarães Fidencio[1], Flávio Ceci[1,2], Alexandre Leopoldo Gonçalves[2,3]

[1]Curso de Ciência da Computação e Sistemas de Informação. Universidade do Sul de Santa Catarina (UNISUL) – Palhoça – SC – Brasil

[2]Departamento de Engenharia e Gestão do Conhecimento – Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC – Brasil

[3]Campus Araranguá – Universidade Federal de Santa Catarina (UFSC) – Araranguá – SC - Brasil

vini.roggia@hotmail.com, paulo.phgf@gmail.com, flavio.ceci@unisul.br
alexandre.goncalves@ararangua.ufsc.br

Resumo. *Atualmente, a rápida expansão na geração de informações em meio digital, seja na Web ou nas organizações em geral, promove desafios, principalmente em como lidar com essa informação que se apresenta de diferentes maneiras e formatos. Por outro lado, cria oportunidades na implementação de soluções que consigam, de maneira rápida e eficiente, encontrar a informação adequada para determinada demanda. Esse artigo propõe um sistema de recuperação de informação que realize consultas em documentos textuais e através da expansão dessas consultas, obtidas através de anotações das palavras-chave contidas em cada documento, possibilite a obtenção de conteúdo relacionado, tais como imagens, vídeos e outros sites similares. Para validar o sistema foi desenvolvido um protótipo e aplicado um estudo de caso em que, professores puderam publicar e anotar determinado documento de uma disciplina visando fornecer aos alunos uma ferramenta para localizar conteúdo relacionado. Uma avaliação inicial realizada por meio de questionários aplicados a um conjunto de usuários mostrou aderência entre a consulta inicial e o conteúdo relacionado obtido como resultado.*

1. Introdução

Tendo em vista o crescimento de dados de diferentes tipos no meio digital e o grande problema para se encontrar informações de nosso interesse, dentro desse meio, - é necessário o uso de tecnologias que facilitem uma busca mais inteligente, trazendo resultados corretos ou que se relacionem diretamente com o termo da busca. De acordo com Santarém e Vidotti (2011), a Recuperação de Informação (RI) tem sido muito discutida na Ciência da Informação, e a busca por informação de qualidade com a necessidade do usuário se tornou pesquisa constante.

Rijsbergen (1979, apud SEIBEL 2007) e Yates et al. (1999, apud SEIBEL 2007) afirmam que "os métodos de recuperação de informação tradicionais se baseiam essencialmente na contagem da frequência em que as palavras aparecem em um documento; sem apresentar soluções para que o conteúdo semântico do discurso seja interpretado". Por não processarem, adequadamente, o documento podem-se perder

importantes informações. Segundo Wives e Loh (1998), muitas das informações disponíveis para acesso rápido e fácil não estão em formatos que possam ser tratados por meios computacionais (imagens, textos, vídeos, gráficos). Batista e Schwabe (2009) asseveram que é difícil o desenvolvimento de aplicações que necessitam capturar e manipular informações diretamente do conteúdo digital, como, por exemplo: a busca de vídeos, sem utilizar seus metadados descritivos (trechos de informação textual associado aos vídeos).

Soma-se a este contexto a *Web* 2.0. A *Web* 2.0 é conhecida como a *web* colaborativa, que surgiu no momento em que as pessoas pararam de apenas consumir informações na *Internet* para também contribuir no processo de elaboração e disseminação de novos conteúdos. Analisando esse cenário verifica-se que o problema de encontrar informações relevantes constitui-se em desafio, tendo em vista a dinamicidade e o aumento crescente do conteúdo gerado.

O objetivo desse artigo é apresentar um sistema que: a) permita o armazenamento de documentos textuais; b) possibilite a recuperação dos documentos através de uma consulta utilizando palavras-chaves; e c) possibilite a expansão dos resultados da consulta utilizando recursos da *Web* 2.0, assim como adicionando conteúdos multimídia, tais como vídeos e imagens, que se relacionem ao resultado obtido. O diferencial desse sistema em relação a outros sistemas de busca será, além da expansão de resultados mostrando conteúdos multimídia, também a possibilidade de adicionar *tags* (palavras-chaves ou etiquetas) a cada documento que for armazenado, permitindo assim, a expansão da consulta através dessas *tags*.

Para concretizar o objetivo foi criado um protótipo que permite o cadastro de documentos e recuperação dos mesmos, trazendo além desses, vídeos, imagens e links para sites, sendo que a expansão dos resultados se baseia nas *tags* dos primeiros documentos encontrados. Essas *tags* são vinculadas ao documento no momento do cadastro de forma manual.

Por fim, é apresentada a validação do protótipo a partir de um estudo de caso, onde professores cadastram os documentos no sistema de modo que o mesmo, seja uma ferramenta para auxiliar no cenário da educação, fazendo com que os professores possam mostrar outros conteúdos relacionados com o tema da disciplina de uma forma rápida e simples. Além disso, tem ainda o propósito de beneficiar os alunos que queiram se aprofundar em determinado tema ou encontrar exemplos diversificados sobre um determinado assunto. Também foi realizada uma pesquisa envolvendo um questionário, onde usuários de diferentes perfis analisaram o sistema proposto.

2. Referencial Teórico

A seguinte seção descreve alguns elementos necessários para o entendimento deste artigo. São abordados conceitos básicos relacionados à Recuperação de Informação (RI), Extração de informação (EI), Expansão de consulta e *Web* 2.0.

2.1. Recuperação de Informação

Conforme Leite (2009, p.7): “o escopo da recuperação de informação pode considerar os recursos de uma área de conhecimento específica como, por exemplo, agricultura, artes, leis e saúde, até os recursos de toda a WWW (*World Wide Web*)”. Para Beppler (2002, p.11) a idéia da RI é simples e objetiva: tendo um armazenamento

de informações relevantes, devemos recuperar apenas a informação desejada, todavia o foco é a maneira como será feito o armazenamento, visto que existem muitos documentos que são difíceis de serem armazenados de forma correta, causando assim, problemas de interpretação no formato digital. Um exemplo disso são os textos em linguagem natural.

Tendo em vista o contexto de *Web*, a EI surgiu como uma forma de aprimorar e organizar os resultados oferecidos pela Recuperação de Informações, extraindo as informações relevantes de algum contexto. (SILVA, 2003, p.2).

Não se pode esquecer que Extração de Informação e Recuperação de Informação possuem conceitos distintos. Ao invés de extrair a informação, o objetivo da RI é selecionar um subconjunto relevante de documentos de uma maior coleção de dados, com base em uma consulta realizada por um usuário (EIKVIL, 1999, p.5). Enquanto a EI é utilizada para filtrar esse resultado da consulta realizada pela RI (SILVA, 2003, p.23). Para realizar a tarefa de extrair as informações de um documento, a EI utiliza métodos que, geralmente, envolvem a escrita de códigos, conhecidos como *wrappers*, que mapeiam o documento para algum modelo de representação do conhecimento. (MARINHO, 2003).

2.2. Expansão de Consulta

Segundo Xu (1996), um problema fundamental na RI é que os autores nem sempre usam as mesmas palavras que os usuários para descrever o mesmo conceito.

A expansão de consulta é uma técnica na qual se busca aumentar a quantidade de termos que devem ser buscados, sendo que estes devem possuir certo grau de equivalência entre si para aumentar a probabilidade de se encontrarem documentos relevantes (CARDOSO, 2002). Para Leite (2009) "a expansão de consulta consiste em adicionar novos termos semanticamente relacionados como os termos presentes na consulta inicial em função do conhecimento contido em uma base de conhecimento [...]".

Segundo Cardoso (2002) os vários métodos de expansão de consulta podem ser divididos em dois principais grupos: métodos iterativos e métodos automáticos.

Métodos iterativos, *User FeedBack Relevance*, também, conhecidos por expansão semi-automática é, provavelmente, a técnica mais comum de expansão de consulta. Conforme afirma Bettio (2007) essa técnica requer que o usuário atribua relevância a um conjunto de documentos trazidos através de uma busca inicial, ou seja, é necessário interação do usuário com o sistema.

Segundo Christopher (2008) os métodos automáticos, diferentes dos iterativos não necessitam de interação com o usuário, o que os torna uma técnica mais interessante, uma vez que o processo é transparente para o usuário.

Conforme afirma Fernandes (2010), o mecanismo de expansão de consultas é essencial no processo de recuperação da informação. Com a ajuda desse, as pesquisas nessa área permanecem com abordagens mais voltadas para o usuário, que possui um papel central nesse contexto.

2.3. Web 2.0

O termo *Web* 2.0 surgiu em 2004, durante uma conferência promovida pelas empresas de mídia Media-Live e O'Reilly Media. Nessa conferência, discutiu-se a idéia de que a Web deveria ser mais do que apenas uma plataforma, deveria ser dinâmica e interativa e colocar o usuário no centro disto.

De acordo com O'Reilly (2005):

> Não há como delimitar fronteiras para a *web* 2.0, pois trata-se de princípios e práticas para que diversos sites sigam. Um dos princípios fundamentais é a *web* como plataforma, ou seja, o usuário poder realizar atividades online que antes só eram possíveis com programas rodando em seu computador.

Pita e Paixão (2010) afirmam que, na *Web* 2.0, os usuários tomam um papel mais ativo, publicando conteúdo, ao invés de apenas consumir. Um exemplo das aplicações web, que ajudaram a construir a *Web* 2.0, que conhecemos hoje, são as redes sociais.

Coutinho e Bottentuit (2007) concluem que a *Web* 2.0 é uma forma de utilização colaborativa da *Internet*, em que o conhecimento é compartilhado de maneira coletiva e descentralizado de autoridade para utilizá-lo e reeditá-lo.

3. Arquitetura Lógica e Modelo

A figura 1 ilustra a arquitetura lógica do sistema proposto, envolvendo um cenário acadêmico, onde são armazenados artigos para, posteriormente, serem recuperados pelo sistema, contudo retornando outros tipos de arquivos relacionados ao artigo, como vídeos, textos ou imagens.

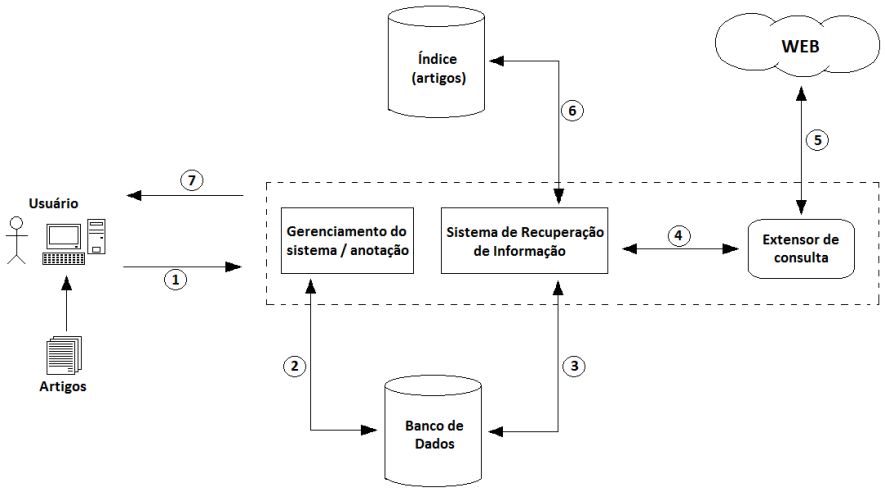

Figura 1. Arquitetura do sistema proposto

A arquitetura do sistema, como exibido na imagem acima, é formada basicamente das seguintes camadas:

a) Camada de apresentação: é a interface *web*, a qual o usuário irá interagir, sendo responsável pelo envio (1) das informações ao sistema, como o cadastro dos dados do usuário e os documentos (artigos) que serão gravados no banco. Através dessa camada, o usuário também poderá visualizar as informações recuperadas (7).

b) Camada de negócio: é o núcleo do sistema, sendo formado pelos módulos de gerenciamento do sistema, sistema de recuperação de informação e extensão de consulta. O gerenciamento do sistema irá gerenciar os usuários e realizar o cadastro dos documentos. O Sistema de Recuperação de Informação (SRI) deve indexar os documentos cadastrados e recuperá-los através do termo de busca. E a extensão irá trazer informações multimídia da *web*, como vídeos, imagens e textos.

c) Camada de dados: é representada pelo banco de dados e informações retornadas da *web*. O banco de dados irá armazenar as informações cadastradas pelo usuário (2), além de armazenar os documentos que irão ser indexados pelo SRI (3). As informações retornadas da *web* (5) serão mostradas ao usuário (7), através do Extensor de consulta (4) que irá realizar a busca em alguns sites pré-determinados.

4. Proposta de solução

Esta seção irá mostrar a arquitetura física do sistema de uma forma um pouco mais detalhada. A Figura 2 ilustra a arquitetura como um todo; ela, entretanto, é dividida em 3 módulos: gerenciador do sistema, indexação e recuperação e extensão de consulta que serão descritos nos próximos tópicos dessa seção.

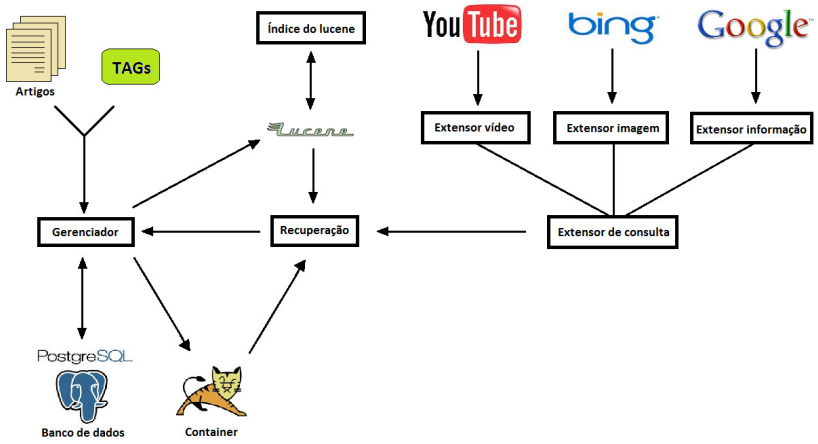

Figura 2. Arquitetura física do sistema

4.1. Gerenciamento do sistema

Este módulo é responsável pelo gerenciamento de informações no sistema, permitindo que o usuário adicione o documento (artigo) para a indexação, além das *tags* referentes ao artigo, esta etapa é a mesma indicada pelo número 1, na Figura 1. Após o usuário realizar a operação de upload do documento desejado e informar as *tags* para esse documento, o sistema irá salvá-lo em um diretório dentro do container *web* (tomcat) e irá gravar os dados referentes a esse documento, como o nome e o caminho, onde foi salvo, no banco de dados (postgres). Ao salvar as informações no banco, o ID (*Identity Document*) desse registro e o próprio texto do documento são recuperados para, posteriormente, serem utilizados na indexação pelo sistema Lucene.

Outra função desse módulo é a visualização das informações do documento. Após uma busca, os resultados são apresentados na tela de resultados. Esse é o passo de número 7, na Figura 1.

4.2. Indexação e recuperação

Um dos fatores para a escolha do SRI Lucene foi ele ter sido desenvolvido com base na linguagem Java, o que oferece uma maior flexibilidade na implementação do protótipo de solução, que é desenvolvido nesta linguagem. Outro fator para a escolha do Lucene é que, de acordo com Hatcher e Gospodnetic (2005), essa é a biblioteca de recuperação de informação mais popular entre as existentes.

Segundo Apache Software Foundation (2011), o projeto Apache Lucene desenvolve software open-source de pesquisa, incluindo Apache Lucene Core, antigamente chamado de Lucene Java, - este fornece indexação de dados e implementações de pesquisa com base na linguagem Java.

Gospodnetic e Hatcher (2005) afirmam que o Lucene é uma biblioteca de recuperação de informação de alto desempenho, projetada para ser agregada a sistemas de indexação e pesquisas textuais em acervos de documentos eletrônicos.

A função desse módulo é indexar o documento, utilizando a ferramenta Lucene. Após o documento ser salvo no banco de dados (postgres), o texto, as *tags* (palavras-chaves) e o ID do registro formam o índice, que será utilizado e gravado pelo Lucene. Essa etapa pode ser relacionada com o passo 2 e 6 da Figura 1.

Para a recuperação do conteúdo, o Lucene irá retornar às informações contidas no índice com base no termo informado pelo usuário e com base nas *tags* que estão gravadas no índice. Assim sendo, com os IDs retornados poderão ser recuperadas as informações gravadas no banco, como também o próprio arquivo gravado no container (tomcat). Esse passo é o mesmo indicado pelo número 3, na Figura 1.

4.3. Expansão de consulta

Quando o usuário informa um termo para a busca, o sistema recupera alguns artigos relacionados a esse termo através das *tags* do documento que, no caso, também, irão servir para recuperar outros tipos de arquivos da *web*. Podemos visualizar na figura 3 como ocorre a recuperação das informações multimídia.

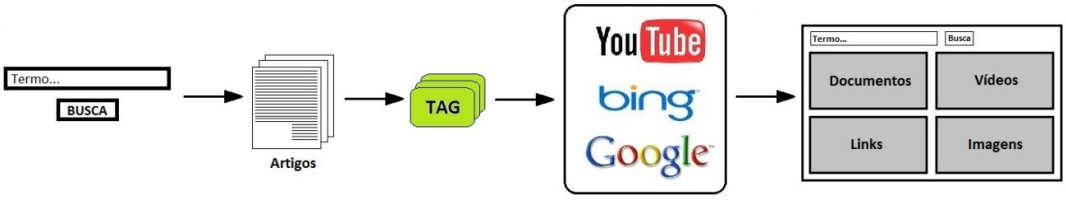

Figura 3. Extensão de consulta

O extensor de consulta por imagem recupera algumas imagens, utilizando a API (*Application Programming Interface*) do Bing, um motor de pesquisa da Microsoft. O extensor de consulta, por vídeo, irá trazer *links* de alguns vídeos do YouTube® para visualização na página de resultados. E o extensor de consultas, por texto, irá apresentar os primeiros resultados do sistema de busca Google, lembrando que todas as informações recuperadas estão diretamente relacionadas às *tags* dos documentos recuperados.

4.3. Estudo de caso

Para validar o sistema foi desenvolvido um protótipo com base em um estudo de caso, sendo que esse gira em torno de um cenário acadêmico, onde os professores poderão publicar, no sistema, artigos de interesse da comunidade acadêmica e os alunos poderão buscar os artigos publicados, podendo também encontrar outras informações que se relacionem com os resultados encontrados.

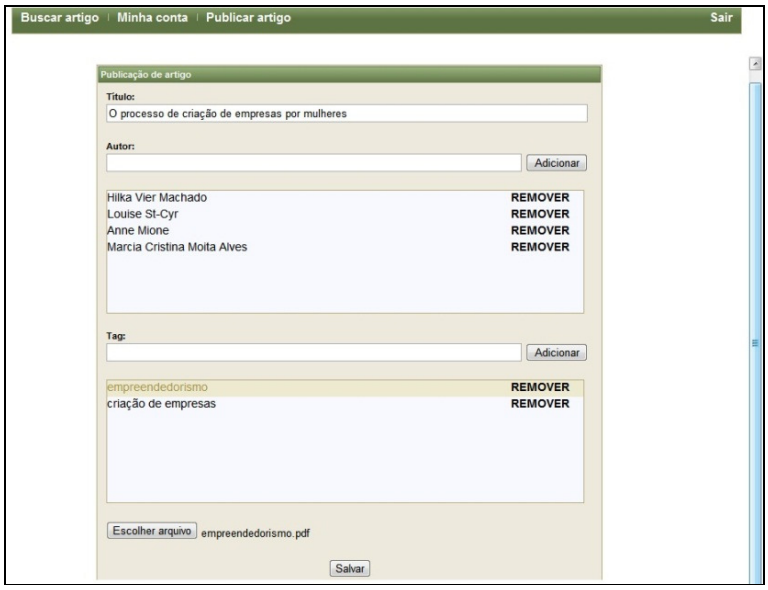

Figura 4. Tela de publicação

A figura 4 ilustra a tela de publicação, onde o usuário de perfil professor poderá cadastrar os artigos no sistema e também adicionar *tags* que se relacionem com o assunto abordado pelo artigo.

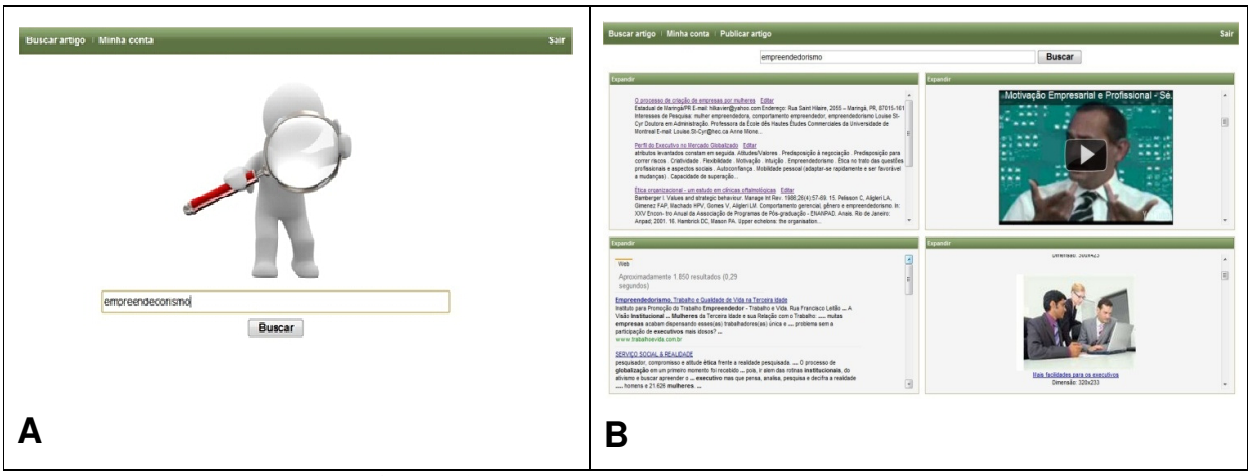

Figura 5. (A) Tela de busca - (B) Tela de resultado

Na figura 5 (A), pode-se observar a tela onde os usuários poderão efetuar a busca informando um termo. O sistema recupera os artigos que possuírem em seu conteúdo o termo informado. Esses artigos recuperados, então, são listados na tela de

resultados (Figura 5, B), junto com as informações recuperadas pela extensão de consulta, lembrando que esta recupera os dados através das *tags* cadastradas nos artigos.

Outra forma de validação do sistema foi através de um questionário com 10 questões, levando em consideração a qualidade dos resultados obtidos pela busca, além da facilidade e desempenho do sistema. A amostra foi composta por 10 entrevistados, de ambos os sexos, com idades entre 19 e 45 anos. Participaram da pesquisa profissionais de diferentes áreas, como tecnologia da informação, direito, administração, psicologia e comércio. A tabela 1 exibe o resultado da pesquisa e as questões abordadas.

Analisando o sistema proposto, podemos então imaginar um cenário, onde professores possam dar aulas utilizando este sistema de busca. O professor, então, pode mostrar a seus alunos artigos de um determinado tema, como também mostrar vídeos, imagens e outros *links* que se relacionem com o assunto dos artigos recuperados, fazendo com que a aula se torne um pouco mais produtiva e interessante. O aluno pode usufruir do sistema para encontrar mais informações referentes ao tema da aula, assim como poderá visualizar dados, que o auxiliem na compreensão de um determinado assunto (imagens e vídeos).

Questão / Resultado	Atende completamente	Atende	Atende em partes	Não atende
Efetua o registro de documentos/artigos científicos e indexa estes para posterior recuperação.	9	1	0	0
Recupera informações de documentos/artigos previamente registrados no sistema.	9	1	0	0
O sistema traz resultados multimídias relacionados à busca do documento/artigo e seus resultados.	6	4	0	0
O sistema traz resultados relevantes aos termos de busca.	8	2	0	0
Tem desempenho satisfatório quanto ao tempo de busca.	6	1	3	0
O sistema tem interface amigável, ou seja, é fácil de manuseá-lo.	6	2	2	0
Apresenta uma forma interessante de exibir resultados.	6	3	1	0
A solução apresentada neste sistema facilita encontrar informações pertinentes aos termos de busca.	8	2	0	0
Permite criar uma base de dados de documentos/artigos de fácil manutenção.	8	2	0	0
Em uma pesquisa simples no sistema, geralmente, as informações desejadas são encontradas.	8	2	0	0

Tabela 1 - Resultado da pesquisa

5. Conclusão

Com relação ao sistema desenvolvido, pode-se afirmar que as APIs utilizadas para a extensão de consulta, possuem algumas limitações perante à busca. O caso mais

relevante é da API do YouTube®, sendo que a base de dados não é muito abrangente, restringindo uma boa parte dos resultados de uma determinada *tag*. Já as consultas realizadas no Bing Imagens e no Google trazem resultados satisfatórios, com relação às *tags* relacionadas ao documento.

Analisando os problemas apresentados no início deste trabalho, a proposta de solução, e os resultados obtidos através da pesquisa, pode-se concluir que a proposta é válida, podendo ser usada em outros cenários e não apenas no apresentado no estudo de caso. A proposta se apresenta como uma alternativa aos buscadores atuais, trazendo vários tipos de dados relacionados com os resultados obtidos.

Pode-se concluir também que a técnica de Extensão de Consulta, utilizada no sistema desenvolvido, mostrou-se uma técnica interessante, mesmo não sendo utilizada pelos buscadores mais conhecidos atualmente. A extensão de consulta de certa forma enriquece as informações retornadas de uma busca.

É valido afirmar que o sistema entra no escopo da *Web* 2.0, porque permite que os usuários compartilhem informações, de uma maneira mais fácil, além de permitir a visualização de mais de um tipo de dado, em uma mesma tela de forma clara e organizada.

Com base nos resultados da pesquisa, pode-se, então, concluir que o sistema atende, completamente, à maior parte de seus objetivos. O sistema recupera corretamente um artigo que foi publicado, sendo utilizadas palavras que estejam no corpo do artigo como termo para a pesquisa. Também traz vídeos, imagens e outros *links* relacionados às *tags* pertencentes aos artigos retornados pela busca. Entretanto o sistema não possui um desempenho satisfatório com relação ao tempo de busca. Em alguns casos, não são retornados vídeos no resultado, devido à limitação do próprio YouTube® referente às informações que o mesmo possui em sua base de dados.

Como trabalhos futuros vislumbram-se melhorias nas anotações dos documentos indexados, através de técnicas de anotação automática ou semiautomática, evitando *tags* que não se relacionem com o conteúdo do documento. Adicionalmente, a utilização de ontologias para suportar buscas semânticas, como uma importante ferramenta, que agregue valor ao processo de recuperação de informação. Por fim, vislumbra-se o desenvolvimento de um coletor de informações (*web crawler*) que leve em consideração o contexto de determinado domínio, visando contornar assim o problema das APIs utilizadas na extensão de consulta.

6. Referencias

Batista, Carlos Eduardo C. e Schwabe, Daniel. (2009) “LinkedTube: Informações Semânticas em Objetos de Mídia da Internet.”, In: Simpósio Brasileiro de Sistemas Multimídia e Web, Fortaleza.

Beppler, Fabiano Duarte. (2002) “Emprego de RBC para recuperação inteligente de informações”, Universidade Federal de Santa Catarina, Florianópolis.

Bettio, Raphael Winckler. (2007) “Interrelação das Técnicas Term Extraction e Query Expansion Aplicadas na Recuperação de Documentos Textuais”, Universidade Federal de Santa Catarina, Florianópolis.

Cardoso, O. N. P. (2002) “Recuperação de Informação”, Departamento de Ciência da Computação – Universidade Federal de Lavras, Lavras, p. 6.

Christopher, D. M., Raghavan, R. e Schütze, H. (2008) “Introduction to Information Retrieval”, Cambridge University Press.

Coutinho, Clara Pereira., Bottentuit Junior, João Batista. (2007) “Blog e Wiki: os futuros professores e as ferramentas da web 2.0”, In: Simpósio Internacional de Informática Educativa, Portugal, p. 199-204.

Eikvil, Line. (2011) “Information Extraction from World Wide Web: A Survey”, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1. 1.41.4905&rank=2, Julho.

Fernandes, Ricardo Madeira. (2010) “GeoSen_Tags**:** um motor de busca geográfico com suporte a Tags”, Universidade Federal de Campina Grande, Campina Grande.

Hatcher, Erik. e Gospodnetic, Otis. Lucene in action. Greenwich: Manning Publications, 2005.

Leite, Maria Angelica de Andrade. (2009) “Modelo Fuzzy para Recuperação de Informação Utilizando Múltiplas Ontologias Relacionadas”, Universidade Estadual de Campinas, Campinas.

Lucene Project (2011). Página do projeto Lucene, http://lucene.apache.org, visualizado em 20/09/2011.

Marinho, Leandro Balby e Giraldi Rosario. (2003) “Mineração na Web”, Sociedade Brasileira de Computação: Revista Eletrônica de Inicialização Científica, http://143.54.31.10/reic/edicoes/2003e2/, junho.

O'reilly, T. (2005) “What is web 2.0: design patterns and business models for the next generation of software”, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/ what-is-web-20.html# mememap

Pita, Marcelo. e Paixão, Goedson Teixeira. (2010) "Arquitetura de Busca Semântica para Governo Eletrônico". In: II Workshop de Computação Aplicada em Governo Eletrônico & Congresso da Sociedade Brasileira de Computação, Belo Horizonte.

Rijsbergen, Van C. J. (1979) “Information Retrieval”, London: Butterworths.

Santarém, José Eduardo S. e Vidotti, Silvana Aparecida Borsetti G. (2011) “Representação iterativa e folksonomia assistida para repositórios digitais”, http://revista.ibict.br/liinc/index.php/liinc/article/view/414/294, março.

Silva, Tércio de Morais Sampaio. (2003) “Extração de Informação para Busca Semântica na Web Baseada em Ontologias”, Universidade Federal de Santa Catarina, Florianópolis.

Xu, J. e Croft, W. B. (1996) “Query expansion using local and global document analysis”, In: SIGIR ’96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, p. 4–11.

Wives, Leandro K. e Loh, Stanley. (1998) “Recuperação de Informações usando a Expansão Semântica e a Lógica Difusa”, In: Congresso Internacional em Ingenieria Informatica, Buenos Aires.