UNIVERSIDADE CATÓLICA DE PELOTAS
ESCOLA DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Um Sistema de Recuperação de Artigos Científicos Baseado em Consultas por Exemplo

por

Christiano Martino Otero Avila

Dissertação apresentada como
requisito parcial para a obtenção do grau de
Mestre em Ciência da Computação

Orientador: Prof. Dr. Stanley Loh

DM-2008/1-003

Pelotas, fevereiro de 2008.

SUMÁRIO

Lista de Abreviaturas

EPC-P	Extrator de Palavras-chave por Freqüência de Padrões
EPC-R	Extrator de Palavras-chave por Freqüência de Radicais
IDF	Inverse document frequency
QBE	*Query by Example*
SAPU	Sistema de Apoio da UCPEL
SISRECAC	Sistema de Recuperação de Artigos Científicos
TF	Term-frequency

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho apresenta um sistema de recuperação de artigos científicos, denominado SisRecAC, que tem por objetivo encontrar na Web, através de uma meta-busca, artigos científicos que sejam de interesse de um usuário. O interesse do usuário é dado através de um texto exemplo; por isso, o sistema é considerado como sendo baseado em consultas por exemplo (*Query by Example*). Uma Consulta por Exemplo é a possibilidade de recuperação de informações a partir de um texto exemplo informado pelo usuário e não a partir de palavras-chave, o que é, atualmente, o padrão para busca de informações. No caso do SisRecAc, informações (palavras-chave) são extraídas do texto exemplo e submetidas ao Google Acadêmico, com isso eliminando a necessidade de formalização, por parte do usuário, da sua necessidade de informação.

A abordagem de Consulta por Exemplo se contrapõe a consultas por palavras-chave e a sistemas de recomendação que utilizam perfis de usuários, na qual o sistema precisa conhecer ou aprender o perfil do usuário e, baseado nessas informações, recomendar ou filtrar objetos relevantes (que coincidam com o perfil do usuário).

Para o desenvolvimento do sistema, foram investigadas técnicas de extração de características de textos, tais como descoberta de termos ou expressões freqüentes no texto, utilização do título do documento, das tags informadas pelo usuário e da combinação dessas técnicas de extração com expansão semântica baseada em Folksonomia. Os métodos, em um total de vinte, foram implementados no sistema e avaliados sob uma ótica subjetiva, questionando o usuário sobre a relevância dos artigos recomendados, e também sob uma ótica mais objetiva ou matemática, avaliando-se a similaridade dos documentos recomendados pela ferramenta em relação ao texto exemplo.

Palavras-Chave: Extração de palavras-chave, sistemas de recomendação, query by example, artigos científicos, meta-busca, expansão semântica, Folksonomias

Abstract

This work presents a recommender system for scientific papers, called SisRecAC. Its goal is to find scientific papers in the Web that may be interesting to the user. The user´s interest is given through an example-text. For this reason, the system is under a Query By Example (QBE) paradigm. Contrary to the approach where the user gives keywords for information retrieval, a system that utilizes QBE receives a text as example of the user´s interest. In the case of the SisRecAC, keywords are extracted from the example-text and then given as input to a Web search engine, using a meta-search approach (in the case, the search engine is the Brazilian Academic/Scholar Google). That way, the user does not need to formulate a query with keywords.

By other side, the QBE approach does not have to create a user´s profile as in filtering or traditional recommender systems.

The work investigated several methods for extracting keywords and characteristics from texts, as for example discovery of frequent single terms and expressions, terms from the title or tags associated to the text by the user. Furthermore, the work utilized combinations of these techniques and also employed semantic expansion on the words extracted from the text. Semantic expansion makes use of Folksonomies and social Web sites. All these methods are described in this dissertation and experiments comparing the methods are discussed. Two kinds of evaluations were performed: one subjective, using human judges, and other objective, evaluating similarity between the example-text and the recommended papers.

Keywords: keyword extraction, recommender systems, query by example, scientific papers, meta-search, semantic expansion, Folksonomy

.

1. Introdução

Atualmente, é grande a quantidade de informações disponível para as pessoas que lêem jornais e revistas, assistem telejornais e ouvem rádio. Porém, todo esse conteúdo é ainda maior para aqueles que, além das mídias tradicionais citadas, acessam regularmente à internet. Essa grande quantidade de informações, que cresce de forma exponencial, faz com que surja um problema de sobrecarga de informação. Apesar do grande volume de informações disponível, muitas vezes as pessoas não conseguem localizar um determinado conteúdo necessário, ou levam muito tempo para localizá-lo.

Especificamente, no caso da Internet, podemos citar os mecanismos de busca, que procuram facilitar a localização de informações. Porém, devido a questões que serão posteriormente analisadas nesse documento, podem acabar gerando mais sobrecarga de informações, e muitas vezes com assuntos irrelevantes para o usuário

Uma das dificuldades encontradas pelos usuários é determinar quais palavras devem ser submetidas ao mecanismo de busca. Conforme Spink et al. (2001), 52% das consultas em mecanismos de busca são reformuladas, sendo o número médio de consultas por sessão de 4,86. Segundo o estudo de Spink et al. (2001), 32,5% das consultas modificadas sofreram alterações nos termos submetidos, mas não no número total de termos; das restantes, 41,6% incluíram termos novos e apenas 25,9% eram relativas a consultas modificadas pela exclusão de termos. A conclusão do referido estudo é que os usuários estão mais inclinados a adicionar do que excluir termos, mas o número de novos termos não é muito grande (5 termos ou menos). Corroborando com estas conclusões, uma pesquisa da iProspect (Iprospect, 2006) concluiu que 82% dos usuários de mecanismos de busca refazem consultas que não foram bem sucedidas, usando o mesmo mecanismo, porém acrescentando mais termos para refinar a busca.

Uma das causas de fracassos na busca, a qual se confirma em nossos estudos (seção 3.7), é o número baixo de palavras submetidas pelos usuários. Segundo diversos estudos, os usuários costumam usar somente de 2 a 3 termos para busca. Silverstein et al. (1999) encontraram uma média de 2,35 palavras por consulta, com desvio padrão de 1,74, e concluíram que apenas 12,6% das consultas utilizam mais de 3 termos. Lau & Horvitz (1999)

encontraram uma média de 2,3 termos por consulta, enquanto que Spink et al. (2001) encontraram uma média de 2,16 termos por consulta. Mais recentemente, Teevan et al. (2006) fizeram um estudo semelhante e encontraram uma média de 2,7 termos por consulta.

Belkin et al. (1997) criaram a Teoria do Estado Anômalo de Conhecimento (ASK - *Anomalous State of Knowledge*). Segundo esses autores, o usuário não está apto a especificar precisamente o que é necessário para resolver essa anomalia, sendo algo irreal pedir ao usuário para formular o que precisa se é isso justamente o que falta. Por esta razão, acredita-se que uma boa estratégia para busca e localização de informações seja um sistema de Recuperação que forneça ao usuário documentos relevantes sem que este precise definir uma consulta ou selecionar palavras-chave.

Uma das alternativas viáveis são os Sistemas de Recomendação ou de Filtragem. Resnick e Varian (1997) explicam que os Sistemas de Recomendação ajudam e incrementam o processo natural e social de auxílio na tomada de decisões do dia a dia. Dizem ainda que, em um sistema de recomendação típico, pessoas fornecem recomendações como entrada, que o sistema então agrega e direciona para a audiência apropriada. Os autores fazem um estudo de alguns sistemas de recomendação e citam o Tapestry (Goldberg et al,1992), como o primeiro sistema de recomendação desenvolvido que se tratava de um sistema experimental de e-mail (*mail system*), desenvolvido no Centro de Pesquisa da Xerox em Palo Alto, cuja motivação era filtrar a grande quantidade de e-mails que começava a incomodar os usuários.

É interessante destacar que foram os desenvolvedores do Tapestry (Goldberg et al,1992) os primeiros a utilizar o termo "Filtragem Colaborativa", assunto que tem sido alvo de muitas pesquisas. Resnick e Varian (1997) preferem utilizar o termo mais geral, "Sistemas de Recomendação" e justificam explicando que as pessoas que recomendam podem não colaborar explicitamente com os destinatários da recomendação, que podem ser desconhecidos e em segundo lugar, recomendações podem sugerir particularmente itens interessantes, além de indicar aqueles que devem ser eliminados.

Sistemas de recomendação têm por objetivo auxiliar no processo social de indicar ou

receber indicação, seja esta indicação referente a livros, artigos, discos, restaurantes ou informações (Resnick & Varian, 1997). Sistemas de recomendação são largamente usados em comércio e marketing para sugerir produtos ou fornecer informações para ajudar o cliente a decidir a compra (Schafer et al, 2001). Nessas aplicações, as pessoas não precisam solicitar ao sistema que forneça as recomendações, mas este decide o que e quando sugerir. No contexto de busca na Web, um sistema de recomendação poderia identificar interesses dos usuários, automaticamente selecionar termos referentes a essas áreas de interesse e submeter tais termos a mecanismos de busca, diminuindo o esforço do usuário e trazendo assim resultados mais relevantes.

Uma outra abordagem para recuperação de informações, e que será melhor descrita no capítulo 2, é a Consulta por Exemplo (*Query by Example*), onde o usuário informa sua necessidade de informação através de um documento exemplo. O sistema, então, extrai características desse documento exemplo e recupera outros documentos similares. Nessa abordagem não é necessário que o usuário saiba expressar sua necessidade de informação através de termos (como nos sistemas tradicionais de busca ou de recuperação de informação) nem tampouco é necessário que o sistema tenha armazenado um perfil deste usuário (como nos sistemas de filtragem ou de recomendação tradicionais).

Neste sentido, em termos gerais, esse trabalho apresenta o SisRecAC (Sistema de Recuperação de Artigos Científicos) que objetiva recomendar artigos científicos de interesse de um usuário. Dito interesse não é informado explicitamente, através de termos ou frases, mas é dado através de um documento exemplo; por isso, o sistema diz ser baseado em consultas por exemplo (Query by Example).

No capítulo 2 são apresentados alguns tópicos importantes da área de sistemas de Recuperação, com destaque para a abordagem de busca a partir de um texto exemplo e extração de características de textos.

No capítulo 3 está descrito o sistema de Recuperação SisRecAC, seu funcionamento na visão dos usuários, funções internas, métodos de extração avaliados, experimentos e resultados.

No capítulo 4 foram relatadas as conclusões e trabalhos futuros.

2. Revisão Bibliográfica

Este capítulo apresenta conceitos e trabalhos importantes das áreas de Recuperação de Informações e Sistemas de Recuperação. Em especial será discutida a estratégia de busca a partir de um texto exemplo (*Query by example* – QBE) e alguns estudos científicos que fornecem subsídios ao desenvolvimento de sistemas que recomendam artigos científicos baseados na extração de palavras-chave de outros documentos. Além destes tópicos, uma seção aborda o tema expansão semântica e o emergente uso de Folksonomias nesta área.

2.1 Consulta por Exemplo (Query by example – QBE)

Na área de recuperação de informações textuais, a idéia de "consulta por exemplo" ou consulta a partir de um texto exemplo, ou o termo consagrado em inglês "*Query by exemple*", remete a uma certa oposição em relação ao padrão atual de busca baseada simplesmente em palavras-chave, tal como o Google (atualmente o mecanismo de busca mais utilizado). Um dos principais problemas desse padrão (por palavras-chave) para busca é a exigência de que o usuário saiba expressar corretamente sua necessidade de informação em termos a serem submetidos ao mecanismo. Esse fato pode ser um problema (Belkin et al., 1997), pois é possível que o usuário não consiga expressar sua necessidade de informação. É importante observar que na medida em que o usuário tem dificuldade em definir as palavras-chave, é possível que ao definir esses termos, consiga apenas em parte, ou seja, defina apenas 2 ou 3 palavras (Silverstein et al.,1999; Lau & Horvitz 1999; Spink et al.,2001; Teevan et al.,2006), fazendo com que o mecanismo de busca recupere uma quantidade muito elevada de textos, sendo necessário, geralmente, selecionar e refinar a busca (geralmente incluindo termos). Tendo esses problemas em vista, o método de consulta por exemplo pode vir a ser uma alternativa interessante em muitos casos.

A tabela 2.1 apresenta 2 paradigmas de sistemas de Recuperação comparados com o paradigma QBE.

Tabela 2.1 – Paradigmas de recuperação de informações.

	Consultas com Palavras-Chave	**Consulta Por Exemplo**	**Filtragem e Recomendação**
Descrição	O usuário deve saber informar corretamente as palavras-chave	O usuário informa um exemplo do que precisa	O sistema constrói um perfil dos usuários (filtragem colaborativa, baseado em conteúdo)
Exemplos	Google, Yahoo, outros	eTBLAST, SisRecAC	Movielens, Grupolens, diversos sistemas de e-commerce
Problemas	Estado anômalo de informação (ASK) – usuário não tem conhecimento suficiente para determinar os termos a serem utilizados	Ter o documento de exemplo	Partida a frio (Cold start

Atualmente, alguns sistemas já implementam a busca por texto exemplo. Em Errami et al (2007) os autores descrevem um sistema de busca chamado eTBLAST, desenvolvido com o objetivo principal de identificar similaridade entre documentos para a descoberta de publicações duplicadas ou plágio. Atualmente, o sistema permite acesso a artigos médicos e biomédicos da base de dados Medline, do Instituto Nacional de Saúde (National Institutes of Health - NIH), do Instituto (Britânico) de Física (Institute of Physics - IOP) e de relatórios técnicos da Agência Espacial Americana (NASA) e está sendo desenvolvido pelo Laboratório de Inovação (The Innovation Laboratory - University of Texas Southwestern Medical School). No eTBLAST, a entrada de dados é baseada na idéia de busca por um texto exemplo (*Query by example* – QBE), pois o usuário pode informar um resumo ou parte de um artigo ou até mesmo todo o artigo (upload) e então obter, como resposta do sistema, uma lista de artigos similares. Nessa listagem, o título de cada artigo é um link que conduz o usuário para uma

página com o resumo do artigo, o nome completo dos autores, o nome da base de dados e um link para o texto completo do artigo.

Um outro importante recurso do sistema é o chamado "*Iterate*", que permite submeter uma nova consulta a partir de um ou mais resumos combinados (mesclados). Para submeter a nova consulta, baseada nos resultados, basta que o usuário escolha um ou mais resultados e clique no botão "*Iterate*".

Um fato a ser destacado em relação ao eTBLAST é que o sistema não armazena permanentemente os documentos enviados pelos usuários e estes não são identificados pelo sistema. As análises são realizadas e ficam disponíveis por um breve período. O armazenamento em caráter permanente e a identificação dos usuários poderiam permitir o envio de avisos sobre novos artigos similares, que eventualmente fossem indexados, ou ainda, possibilitaria a construção de um perfil de cada usuário.

2.2 Extração de Características de Textos

A extração ou identificação de palavras-chave de textos permite a construção de metadados semânticos de textos e se caracteriza por ser uma das formas mais comuns de extração de características. Essa informação pode ser utilizada em uma série de aplicações, dentre elas, destacam-se as que necessitam indexar documentos para pesquisas, ou seja, combinar (*match*) com palavras-chaves digitadas por um usuário que faz uma consulta. Palavras-chave são um conjunto de termos únicos (simples) ou compostos (dupla, trios, etc.) que ajudam a determinar o assunto (ou os assuntos) que são tratados em um texto.

Brooks & Montanez (2006) discutem alguns métodos de extração que utilizam a freqüência das palavras. As mais freqüentes são mais prováveis para serem utilizadas como palavras-chave. Também é utilizado o método TF-IDF, que será melhor explicado neste documento na seção 2.3. Os autores concluem que a geração automática de palavras-chave (mais freqüentes) e o conseqüente agrupamento, produzem grupos mais similares (mais focados) do que os grupos gerados a partir da identificação manual (*tagging*). Os autores destacam também que quando o processo de identificação do texto é manual, as pessoas, muitas vezes, tendem a identificar o texto com termos que possuem um significado pessoal e não identificam o assunto do texto.

Segundo (Pereira et al, 2002), a literatura apresenta diversas técnicas de extração de palavras-chave, porém, em sua maioria, são voltadas à língua inglesa. Com relação especificamente à língua portuguesa, as técnicas descritas são bastante superficiais, como, por exemplo, as baseadas unicamente na freqüência das palavras.

Pereira et al,(2002) analisam métodos de extração baseados na freqüência de padrões léxicos (*lexical compounds*). Seqüências do tipo “nome preposição nome”, “nome adjetivo”, e outras são utilizadas para determinar a relevância de palavras simples ou compostas (duplas ou trios). Além disso, a relevância também é determinada pela posição da palavra no texto; se uma palavra é relevante para o texto, ela deve aparecer na introdução (para os autores, até a posição 450). Já uma palavra que ocorra pela primeira vez após a posição 800 pode ser considerada irrelevante.

Em (Pereira et al,2001) foram apresentados e avaliados dois algoritmos para extração de palavras-chave, o EPC-P (Extrator de Palavras-chave por Freqüência de Padrões) que é baseado em métodos estatísticos para determinar as palavras-chave e o EPC-R (Extrator de Palavras-chave por Freqüência de Radicais) que utiliza o algoritmo Extractor definido em Turney (1999), que foi concebido para o inglês e utiliza apenas a freqüência dos radicais. Para o desenvolvimento do algoritmo EPC-P, que leva em consideração os padrões morfossintáticos encontrados no texto, os autores (Pereira et al,2001) fizeram um levantamento com 58 artigos, escritos em língua portuguesa, da área de Computação. Foram analisadas as palavras-chave definidas pelos autores e os padrões mais freqüentes foram:

- nome;
- nome preposição nome;
- nome adjetivo;
- nome adjetivo adjetivo;
- nome adjetivo preposição nome; e
- nome preposição nome adjetivo.

Posteriormente, os textos tiveram todas as suas palavras identificadas ou etiquetadas, sendo cada uma associada a uma classe gramatical, utilizando o etiquetador (*Part-of-Speech Tagger*) da língua portuguesa (Aires et al., 2000). O próximo passo foi a construção de seis listas, cada uma possuindo as seqüências de texto que se enquadram nos padrões que foram

descobertos com a análise das palavras-chave definidas pelos autores (nome, nome preposição nome, nome adjetivo, nome adjetivo adjetivo; nome adjetivo preposição nome e nome preposição nome adjetivo). Outras seis listas são construídas, uma para cada padrão, porém apenas com os radicais (Porter, 1980) das palavras. Cada uma destas seis listas é ordenada de forma decrescente em relação ao número de ocorrências de cada palavra no texto, já as listas que contém as ocorrências originais permanecem em ordem alfabética.

Após a criação das doze listas, uma lista chamada “lista 1” é criada com os radicais mais freqüentes (freqüência relativa), extraídos das seis listas de radicais que posteriormente são transformados em uma nova lista onde cada radical é convertido para o original correspondente, finalizando com uma lista de trinta palavras-chave para o texto.

Com relação ao algoritmo EPC-R, três listas são construídas, A primeira contém todas as palavras simples do texto; a segunda, todas as duplas; e a terceira, todos os trios de palavras do texto (em ordem alfabética). Em cada lista, além da palavra (simples, em dupla ou em trio), também é armazenado o número de ocorrências da(s) palavra(s) no texto.

Paralelamente, outras três listas são construídas, porém apenas com os radicais das palavras, obtidas também pelo algoritmo de Porter. Quanto à ordenação, as listas com os radicais são ordenadas pela freqüência (decrescente) e as listas com os termos originais são ordenadas alfabeticamente. Posteriormente, a lista 1 é criada com os cinquenta elementos de maior frequência (considerando as três listas de radicais). Posteriormente, é criada uma lista 2, que é baseada na lista de radicais simples (ordem decrescente) e analisando a primeira ocorrência na lista 1 (mais freqüentes das três listas de radicais). Depois desse passo, a lista 2 é utilizada como base para que os elementos sejam transformados em sua forma original, sendo consideradas as trinta palavras de maior ocorrência.

Os autores (Pereira et al,2001) apontam alguns pontos falhos de cada algoritmo, com relação ao EPC-R. Uma das limitações destacadas foi o fato de o algoritmo trabalhar apenas com termos simples, duplas e trios e, portanto, não identifica expressões como “design centrado no usuário” ou “design centrado no aprendiz”. Porém, o EPC-R alcançou *recall* melhor, ou seja, encontrou mais palavras−chave do autor do que o EPC−P. Um outro aspecto negativo destacado pelos autores, com relação ao EPC−R, foi o fato de que foram localizadas muitas palavras sem importância, como “97”, “deve”, “outra”,“pode”, entre outras. O EPC−P, manteve suas palavras−chave mais próximas ao tema central do texto e às palavras−chave do autor pelo fato de o EPC−P estar preso a padrões morfossintáticos,

evitando, por exemplo, considerar relevantes verbos e numerais. Segundo a avaliação dos autores, o EPC-P apresenta um desempenho superior em relação ao EPC-R, gerando palavras com um bom valor representativo.

Em Dias(2004), é apresentada uma adaptação do algoritmo KEA para o Português. Ele utiliza análise léxica e aprendizado de máquina. Um modelo de predição é construído usando textos de treinamento com palavras-chave previamente selecionadas. Tal modelo é usado depois para encontrar palavras-chave em textos novos. Também são analisadas palavras simples ou compostas de tamanho dois ou três. Há algumas regras para eliminar falsos candidatos tais como “palavras candidatas não podem começar ou terminar com uma *stopword* nem podem conter nomes próprios (uma inicial maiúscula)”. Após um processo de *stemming* (redução das palavras a seus radicais), dois graus de relevância são calculados para cada palavra candidata e usados no treinamento e na extração: (a) um grau calculado pelo método TF-IDF, onde a relevância é proporcional à freqüência da palavra num documento e inversamente proporcional à sua freqüência no conjunto de documentos, e (b) um grau relativo à posição da primeira ocorrência da palavra no documento, sendo que este grau é calculado pelo número de palavras que precedem o primeiro aparecimento da palavra, dividido pelo número de palavras no documento. São desconsideradas quaisquer palavras que são sub-palavras de alguma outra com peso de relevância maior. A avaliação do algoritmo foi feita comparando o número de coincidências entre as palavras determinadas pelo KEA e as palavras-chave que foram originalmente escolhidas por autores de documentos.

Em (Pereira et al,2002) também é descrito um trabalho de adaptação do algoritmo KEA (Witten et al, 1999) que permite a extração automática de palavras-chave de textos em português. Foi, ainda, utilizado um radicalizador para a Língua Portuguesa, uma implementação do algoritmo Portuguese Stemmer, proposto por Viviane Orengo e Christian Huyck em Orengo (2001). Em (Pereira et al,2002) são descritos os passos do algoritmo Portuguese Stemmer, que são:

1) Redução do plural;
2) Redução do feminino;
3) Redução do advérbio;
4) Redução do aumentativo e do diminutivo;
5) Redução das formas nominais;

6) Redução das terminações verbais;
7) Redução da vogal temática;
8) Remoção dos acentos;

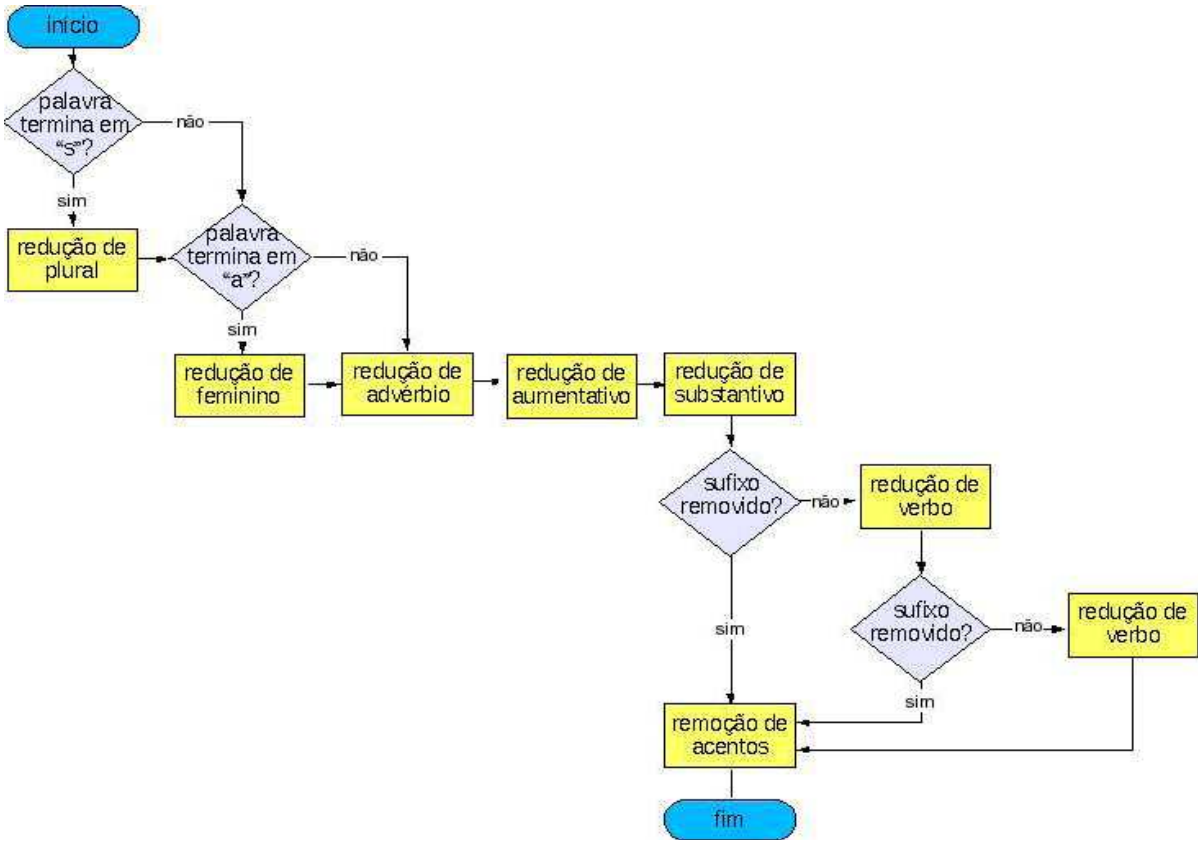

Fig. 2.1 – Passos do algoritmo Portuguese Stemmer (Pereira et al, 2002).

A Figura 2.1 apresenta, em formato de um fluxograma, os passos do algoritmo Portuguese Stemmer.

2.3 TF/IDF

Freqüentemente, em sistemas de recuperação de informação e mais especificamente em sistemas de recomendação, utiliza-se o modelo vetorial definido por Salton (1983) conhecido por tf-idf (*term-frequency inverse document frequency*) . O principal objetivo deste método é identificar termos que, apesar de muito freqüentes em um documento, também aparecem com freqüência em outros textos de outra área. Sendo assim, são considerados de pouco valor discriminatório, pois não ajudam a classificar um texto em áreas.

As fórmulas para o cálculo da medida tf-idf são:

- Freqüência do Termo

A freqüência do termo é dada pela divisão do número de ocorrências do termo dividido pelo número de ocorrências de todos os termos, conforme a fórmula da figura 2.2

$$\mathrm{tf} = \frac{n_i}{\sum_k n_k} \tag{2.1}$$

Fig. 2.2 – freqüência do termo.

- Freqüência Inversa

A freqüência inversa é a medida de importância geral do termo e é o logaritmo da divisão do número total de termos sobre o número de documentos que contém o termo, conforme pode ser visualizado na figura 2.3.

$$\mathrm{idf} = \log \frac{|D|}{|(d_i \supset t_i)|} \tag{2.2}$$

Fig. 2.3 – Freqüência inversa.

- TF/IDF

Portanto a medida tf/idf é obtida pela multiplicação da freqüência do termo pela freqüência inversa, ou seja, tfidf = tf x idf.

2.4 Expansão Semântica

No contexto da busca de informações na web, um dos problemas que os sistemas enfrentam é a ambigüidade dos termos utilizados pelos usuários para localizar informações. Essa ambigüidade faz com que os motores de busca recuperem informações irrelevantes, trazendo documentos de assuntos que não são de interesse do usuário. Outro problema,

apresentando em Belkin et al. (1997), é a dificuldade que os usuários possuem para expressar sua necessidade de informação em termos ou palavras-chave.

Além dessas questões, surge também o problema da não identificação de conceitos que estão relacionados a termos que foram digitados pelos usuários. Por exemplo, se o usuário digitar o termo “MEC”, que é sigla de “MINISTÉRIO DA EDUCAÇÃO”, é possível que esta sigla não esteja explícita em uma página relevante, e, portanto, o mecanismo não recupere essa página, por não identificar o conceito que está associado ao termo “MEC”.

Como forma de resolver essas questões, diversos autores (Lau, Tessa et al,1999; Billerbeck, Bodo et al,2003; Billerbeck, Bodo et al,2004; Fonseca et al,2005; Xu, Jinxi et al,1996; Chekuri et al, 1997; Zeng et al, 2004) desenvolveram estudos no sentido de expandir os termos que foram informados pelo usuário. Alguns estudos como por exemplo, em Chekuri et al, (1997) propõem a utilização de taxonomias existentes nos diretórios Web, como os Yahoo Directories ou ainda a utilização de clusters, como em; Zeng et al (2004) ou ontologias (PARALIC; KOSTIAL, 2003) que viabilizam o cálculo da similaridade entre documentos e os termos informados pelo usuário, mesmo que estes termos não estejam explicitamente no documento.

Ribeiro-Neto et al. (2005) discutem um tema semelhante. Neste caso, o problema deles é relacionar páginas Web com anúncios, ou seja, encontrar anúncios que tratem do mesmo assunto de uma página sendo considerada. Esse problema foi chamado por eles de “*impedance coupling*” ou algo como “acoplamento de impedância”. Os métodos analisados procuram encontrar palavras-chave que estejam ou não nos textos das páginas ou dos anúncios. Para tanto, são comparadas diversas técnicas de expansão semântica.

Sobre o uso de técnicas de expansão semântica para complementar consultas, há alguns trabalhos recentes interessantes. Fonseca et al. (2005) discutem técnicas para expansão semântica de consultas, isto é, para acrescentar novos termos a uma consulta e submetê-la a um mecanismo de busca, com o objetivo de diminuir a ambigüidade. A idéia é gerar um grafo direcionado com relações entre consultas (mineradas através da técnica de associação; duas consultas estão relacionadas se aparecem juntas em outras seções). Um grupo de consultas é considerado um conceito. Os possíveis conceitos são apresentados ao usuário, que então decide qual o melhor conceito para atender sua necessidade. A consulta original então é expandida com termos das consultas relacionadas, dentro do mesmo

conceito.

Os autores Kraft et al. (2006) os autores apresentam três métodos que automaticamente complementam a consulta do usuário com informações capturadas do contexto deste (por exemplo, termos de uma página Web que o usuário está acessando ou um arquivo que o usuário está editando no momento). O principal objetivo do trabalho é determinar o contexto da pesquisa, eliminando ambigüidades que podem surgir nos termos da consulta. Os três métodos apresentados em (Kraft et al, 2006) são:

a) *Query Rewriting - QR*: neste método, um vetor de termos de contexto, que pode ser capturado a partir dos textos abertos no editor de texto do usuário ou páginas web que o usuário está visitando, é concatenado com os termos normais da consulta. Os autores citam a desvantagem de a pesquisa ficar muito restrita, com baixa abrangência ou revocação (recall), tendo em vista o grande número de termos;

b) *Rank Biasing - RB*: este método requer um mecanismo de busca modificado, que receba como entrada os termos do vetor de contexto, um operador (seleção, opcional, etc.) e um peso multiplicador para cada par operador-termo. Um exemplo de consulta estruturada para esse método seria: <selection=cat> <optional=persian, 2.0>. Tal consulta seleciona todos os documentos que contém o termo “cat” e melhora o posicionamento dos documentos que possuem o termo “persian” por um fator definido em 2.0. Esse peso associado determina o quanto o termo deve influenciar o ranking final;

c) *Iterative Filtering Meta-search - IFM*: este método é baseado no conceito de meta-busca e a idéia básica é o envio de múltiplas consultas para o mecanismo de busca. As consultas são reconstruídas com a combinação dos termos da consulta original com os termos do vetor de contexto. Posteriormente os resultados de todas as consultas são unificados e um ranking único é elaborado, utilizando-se uma técnica de agregação de ranking.

A figura 2.4 apresenta uma visão geral dos métodos propostos por (Kraft, R et al, 2006).

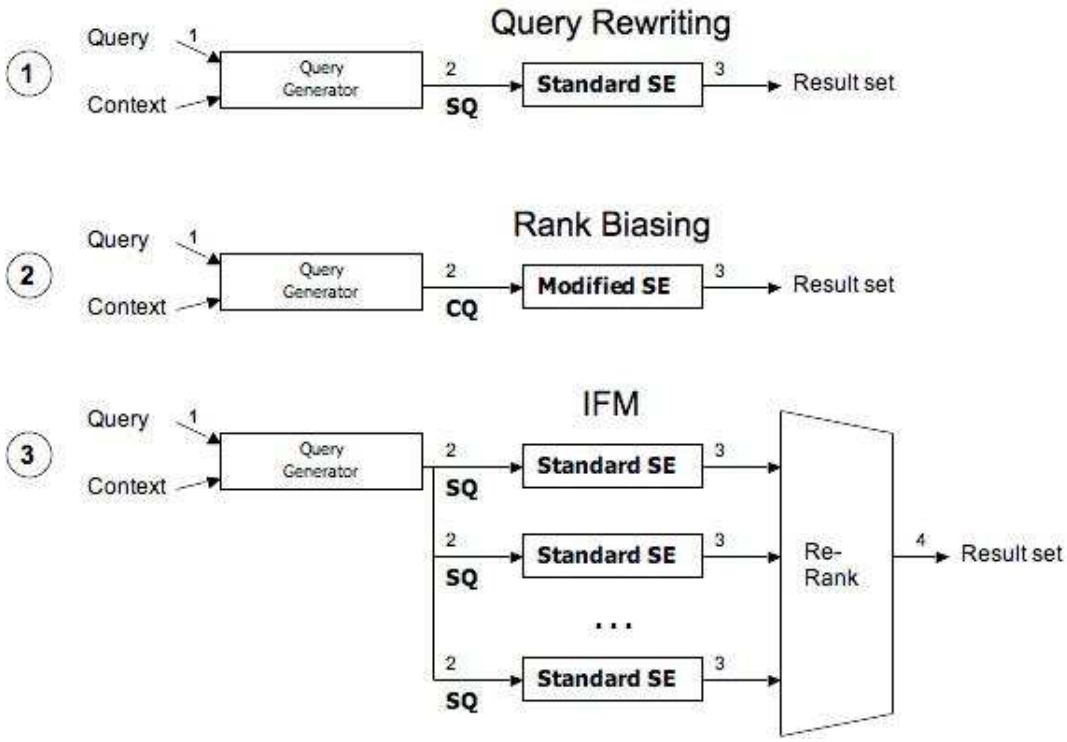

Fig. 2.4 - Métodos propostos por (Kraft, R et al, 2006).

Tabela 2.2. Precisão de diferentes quantidades de termos na consulta(Kraft et al, 2006).

	MSN		Yahoo		Google	
Média de termos na consulta	@1	@3	@1	@3	@1	@3
2,25	50,3%	50,4%	53,1%	49,6%	51,3%	48,9%
4,14	72,8%	68,7%	67,7%	68,8%	72,6%	71,7%
5,73	80,0%	77,0%	78,4%	75,8%	79,3%	78,4%
7,30	79,2%	77,5%	81,3%	80,1%	82,8%	80,1%
8,61	77,5%	75,7%	79,8%	78,0%	81,2%	80,2%

Uma das conclusões interessantes do trabalho de Kraft et al. (2006) é sobre o número de termos utilizados para refazer as consultas (no método *Query Rewriting*). A tabela 2.2 apresenta os resultados de precisão para diferentes médias de termos. A precisão foi avaliada por juízes humanos que decidiam se um resultado era “interessante”, “de algum modo interessante” ou “não interessantes”. Para o cálculo da precisão, foram consideradas corretas as duas primeiras opções (“interessante” + “de algum modo”). Os resultados foram avaliados com três mecanismos de busca (MSN, Yahoo e Google) considerando o 1º

documento recuperado (@1) ou os três primeiros recuperados (@3). Como pode ser notado na tabela 1, os melhores resultados são conseguidos fornecendo de cinco a nove palavras-chave, sendo a melhor média em torno de sete termos. Entretanto, essa conclusão é específica para o método experimentado por Kraft e al. (2006), o qual utiliza um vetor de contexto (e o vetor, para ser elaborado, necessita de informações pessoais do usuário).

Com relação aos métodos, os autores mostram que o QR, que é uma técnica simples de emulação das reformulações de consultas humanas e pode ser facilmente implementada no topo de mecanismos de busca tradicionais, funciona surpreendentemente bem e é provável que seja superior às reformulações humanas. Os métodos RB e IFM, são melhores no aspecto de recuperação em relação ao método QR, sendo que o IFM é muito eficaz e superou os outros dois métodos em termos de recuperação e relevância, mas adiciona um alto custo de engenharia.

Em Ferragina et al (2005), é proposto e implementado um mecanismo de meta-busca chamado SnakeT, que apresenta os resultados em clusters hierárquicos, onde uma abordagem interativa, (Interactive Query Expansion), permite ao usuário escolher o cluster de interesse, permitindo ou facilitando a expansão ou refinamento dos termos da busca. Na fig. 2.5, pode ser observado que a busca pelo termo “jaguar” gerou os clusters “Cars”, “Motors”, “Diego Zoo”, “Search Profile Site”,”Land Rover”,”Largest Big Cat”.

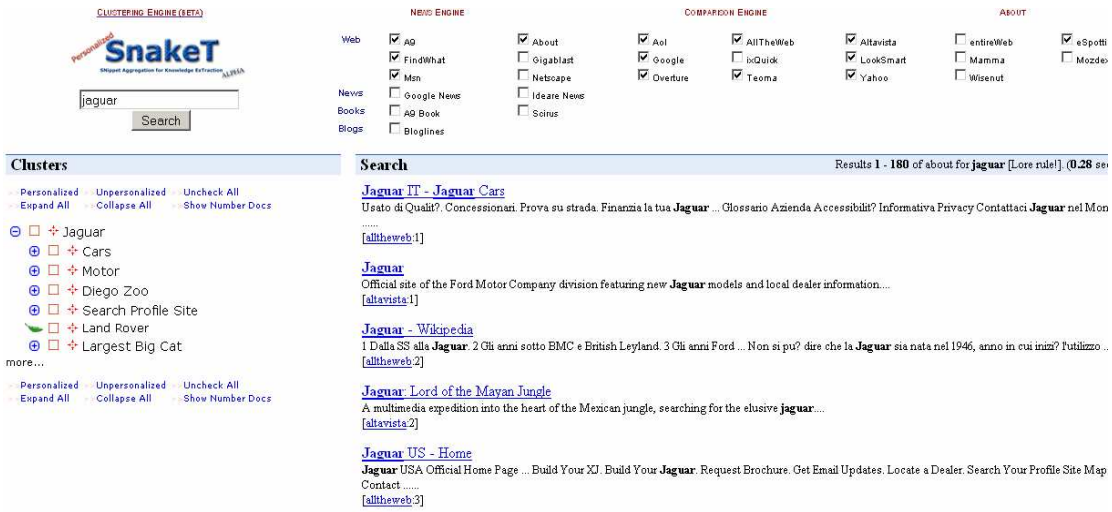

Fig. 2.5 – Mecanismo de busca SnakeT – busca por “jaguar” – Ferragina et al. (2005).

De forma interativa, o usuário pode escolher um dos clusters e com isso expandir ou refinar a sua busca, conforme apresentado na fig. 2.6.

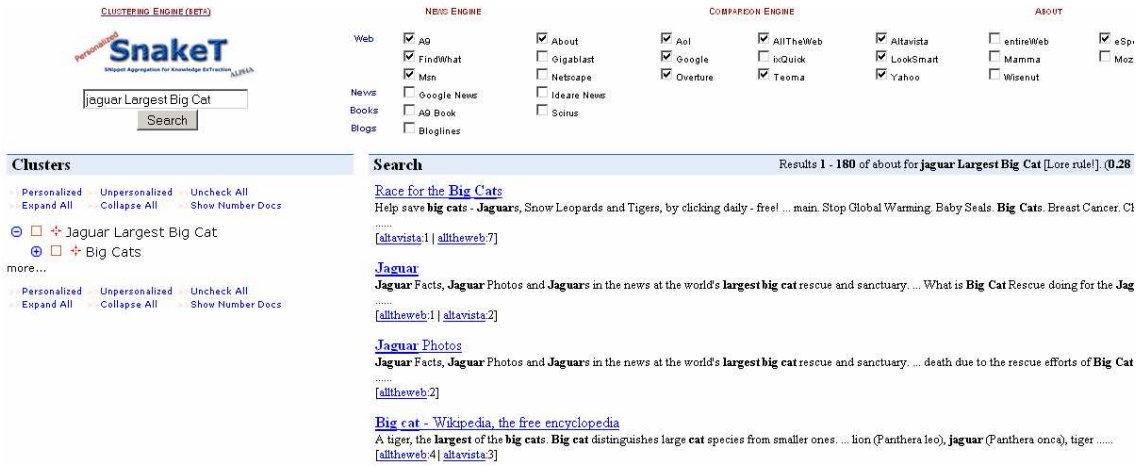

Fig. 2.6 – Mecanismo de busca SnakeT – busca por “Largest Big Cat” – Ferragina et al. (2005).

O SnakeT, por ser um sistema de meta-busca, não possui uma base própria, se utiliza de informações de outros mecanismos de busca, que podem ser selecionados pelo usuário. A fig. 2.7 destaca os sistemas atualmente suportados pelo SnakeT.

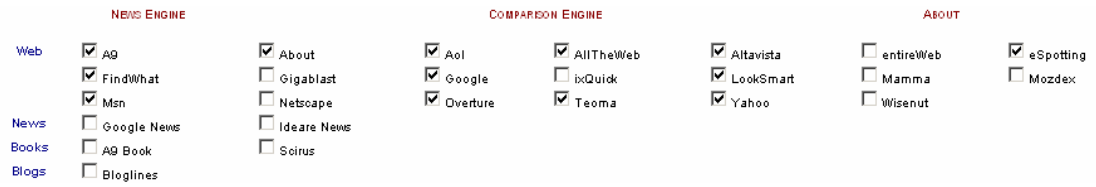

Fig. 2.7 – Mecanismos de busca suportados pelo SnakeT.

Outros exemplos de mecanismos de busca que implementam cluster, e com isso permitem de forma interativa, o refinamento da busca são:

- Clusty[12];

[12] http://clusty.com

- Vivísimo[13];
- Google (experimental)[14];
- Yahoo[15];
- Collarity[16];
- Quintura[17];
- ask[18].

Em (Ribeiro-Neto et al, 2005) são apresentadas algumas técnicas para identificação mais efetiva de anúncios que estejam relacionados a uma determinada página. O autor apresenta dez técnicas para melhorar a associação de anúncios com conteúdos de páginas web, mostrando os anúncios que realmente tenham relação com o conteúdo da página. Segundo o autor, cinco técnicas são baseadas em expansão semântica, ou expansão dos termos da página web para facilitar a tarefa de combinar com os anúncios.

A motivação do trabalho se deu devido à observação que há freqüentemente uma combinação equivocada entre o vocabulário da página web e o vocabulário do anúncio. Em tais casos, o autor afirma que existe um problema na impedância do vocabulário e que suas técnicas provêem efeitos positivos no acoplamento da impedância pela redução do vocabulário de impedância. A proposta básica é acrescentar novos termos (palavras) para a página web para também reduzir o vocabulário de impedância, provendo um segundo

[13] http://vivisimo.com

[14] http://www.google.com/search?hl=pt-BR&esrch=RefinementBarRhsPreview&q=jaguar&btnG=Pesquisar&lr=

[15] http://search.yahoo.com/search?p=jaguar&fr=yfp-t-501&toggle=1&cop=mss&ei=UTF-8&vc=&fp_ip=BR

[16] http://www.collarity.com/

[17] http://www.quintura.com/

[18] http://www.ask.com

caminho de combinação. Os autores se referem à técnica como acoplamento de impedância.

Além da questão dos sinônimos ou identificação de conceitos, surge um outro problema a ser resolvido: a identificação do contexto que um termo está sendo utilizado para eliminar a ambigüidade de uma consulta. Um exemplo clássico desse problema, citado por Kraft, R et al. (2006) é a utilização do termo JAGUAR. Para ilustrar esse problema basta fazer uma busca no Google (figura 2.8), procurando páginas em português, usando o termo JAGUAR. O Google retornou dois anúncios relacionados a automóveis (links patrocinados) e com relação aos dez primeiros links apresentados, seis estão relacionados ao automóvel, dois sobre o animal, dois sobre Força Aérea Brasileira (esquadrão jaguar).

Fig. 2.8 - Problema em relação ao contexto da pesquisa - Jaguar (carro x animal).

2.5 Folksonomias

Folksonomia é um termo criado por um arquiteto da informação chamado Thomas Vander Wal (Smith, 2004) para designar o resultado do processo de identificação de recursos por usuários, através do uso de palavras-chave ou o termo mais utilizado “tag” ou no plural “tags”. O termo em inglês *folksonomy*, advém da junção de *folks* (povo/pessoa) com *taxonomy* (taxonomia). Em síntese, traduzindo para o português, pode-se dizer que seria a "classificação das pessoas ou do povo".

As Folksonomias podem ser encontradas em diversos tipos de sites web que permitem o armazenamento de recursos por parte de seus usuários, tais como o armazenamento de imagens, de vídeos, documentos, músicas e links (URLs) e são também chamados de “*tagging systems*”. A tabela 2.3 apresenta a relação de alguns sistemas que utilizam tags para identificar os recursos armazenados por seus usuários.

Tabela 2.3 – Relação de alguns sistemas que utilizam tags para identificar recursos.

SITE	ENDEREÇO	TIPO DE RECURSO
Del.icio.us	http://del.icio.us	Compartilhamento de links
Bybsonomy	http://www.bibsonomy.org	Compartilhamento de links
Diigo	http://www.diigo.com/	Compartilhamento de links
CiteULike	http://www.citeulike.org	Compartilhamento de links
Flickr	http://www.flickr.com	Compartilhamento de fotos
YouTube	http://www.youtube.com	Compartilhamento de vídeos
Lastfm	http://www.last.fm	Base de dados que permite o compartilhamento de informações sobre músicas, músicos e álbuns.
SlideShare	http://www.slideshare.net	Compartilhamento de arquivos no formato ppt
Scribd	http://www.scribd.com/	Compartilhamento de arquivos em diversos formatos (ppt, pdf, doc, txt,etc...)

Pesquisas recentes têm dedicado atenção ao tema Folksonomia e as possibilidades de uso desse conhecimento. Wu et al. (2006), abordam a descoberta de conhecimento em Folksonomias, geração de ontologias e recomendação de usuários e documentos. Schmitz et al. (2006) discutem como analisar e estruturar Folksonomias e como os resultados podem ser utilizados em ontologias e suporte a semânticas emergentes. Andreas Hotho, Robert J¨aschke, Christoph Schmitz, Gerd Stumme e outros produzem vários artigos na área de Folksonomias e fazem parte de um grupo chamado *Knowledge & Data Engineering Group* of the University of Kassel.

O grupo mantém um sistema chamado Bibsonomy[19], voltado principalmente a pesquisadores que desejam compartilhar *bookmarks* e bibliografias. Eles afirmam oferecer um sofisticado suporte às tarefas de visualização, pesquisa e ranqueamento de *bookmarks* e bibliografias.

[19] http://www.bibsonomy.org

A fig. 2.9 apresenta uma tela de cadastro de um endereço web, também denominado link ou *bookmark* e o local para a informação das tags, com sugestões baseadas no conteúdo da página que está sendo cadastrada e na Folksonomia. A estratégia de sugerir tags, além de auxiliar o usuário na escolha da identificação do recurso, também pode resultar em uma diminuição do número de tags diferentes e idiossincrasias, sendo uma estratégia adotada por outros sistemas que utilizam Folksonomias, pois conduz a um vocabulário um pouco mais unificado (Sood et al.,2007).

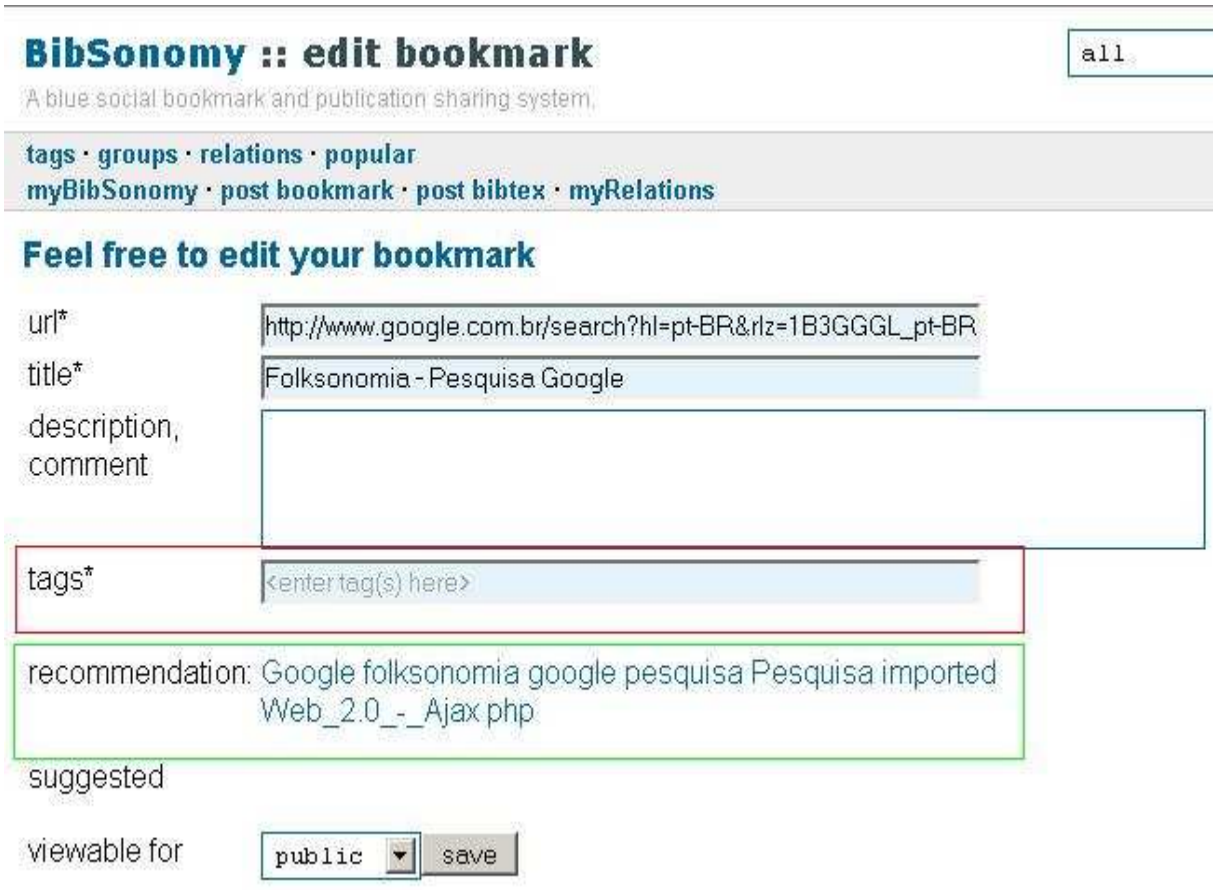

Fig. 2.9 – Bibsonomy – Cadastro de URL (*bookmark).*

Em Begelman et al. (2006), os autores apresentam uma técnica de clusterização de tags para melhorar a experiência do usuário em serviços de *tagging* e conseqüentemente proporcionar um maior sucesso ao serviço. Os autores descrevem o algoritmo que foi utilizado para agrupar (*Clustering Algorithm*) as tags e também algumas técnicas utilizadas para identificar tags relacionadas semanticamente. Informam que foi utilizado o algoritmo de bisecção Espectral (Pothen et al,1990). A técnica para localizar tags relacionadas é

baseada na contagem do número de co-ocorrências, ou seja, de tags que são utilizadas em uma mesma página. Porém não fica claro no artigo como, tecnicamente falando, é realizada a captura dos dados, se por extração diretamente das páginas de resultados de sites de social bookmarking ou mesmo se foi utilizada alguma API desses mesmos tipos de sites.

Os autores explicam, por exemplo, a descoberta tags relacionadas para a tag “RSS’, conforme apresentado na fig. 2.10 . Os autores observam que a tag “RSS” aparece identificando 310 páginas juntamente com a tag “feed”. A mesma tag “RSS” aparece identificando 298 páginas juntamente com a tag “blog”.

tag	**contagem**	*tag*	**contagem**
feed	310	web2.0	77
blog	298	home	65
feeds	246	wikipedia	59
search	219	blogs	57
news	173	biography	53
google	103	preview	48
xml	102	learn	33
web	81	sitemap	30

Fig. 2.10 – Tags relacionadas com RSS (Begelman et al,2006).

Os autores informam também que descobriram um padrão para a distribuição das tags relacionadas e apresentam em um gráfico (fig. 2.11), onde as tags são divididas em fortemente relacionadas e levemente relacionadas.

O eixo Y indica o total de vezes que uma determinada tag (eixo X) aparece relacionada a uma tag, no caso foi usado como exemplo a tag “RSS”.

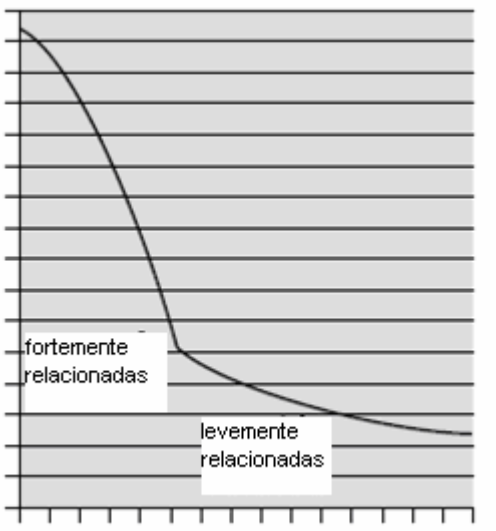

Fig 2.11 – Distribuição padrão de tags relacionadas (Begelman et al,2006).

Tag	tags relacionadas
Apple	mac, osx, macosx, tiger
Art	cool, design, fun, graphics, images
javascript	ajax, dhtml, programming languages
music	audio, media, mp3, ipod, itunes
photography	galleries, photo, hi-res, sexy, flickr, images
software	computers, hardware, acorn, internet, linux, open source software, mambo, programming, technology, web howto, tips, reference, tutorials, tools
free	download, freeware, opensource

Fig. 2.12 – Resultados – Tags relacionadas (Begelman et al,2006).

Os autores apresentam apenas alguns resultados experimentais (fig. 2.12) obtidos com a técnica de descoberta automática de tags relacionadas baseadas em cluster com os dados do RawSugar *tag space* e informam que mais resultados podem ser consultados em http://www.rawsugar.com/lab.

Em Sood et al.(2007) é descrito um sistema chamado TagAssist, o qual fornece sugestões de tags para sistemas de blogs. Segundo os autores, o sistema é capaz de aumentar a qualidade de tags sugeridas. Além disso, o sistema emprega um conjunto de métricas para avaliar a qualidade potencial das tags sugeridas.

Os autores justificam a utilidade do sistema no sentido de reduzir o total de tags para consolidar o vocabulário utilizado, reduzindo tags únicas e idiossincrasias e com isso facilitando a busca de recursos (links, imagens, documentos, músicas e etc) e a navegação

(*browsing*).

O sistema basicamente analisa a nova postagem, que ainda não recebeu as tags de identificação, encontra outras postagens similares, agrega as tags associadas a essa postagem, e então recomenda o conjunto de tags ao usuário da nova postagem. Ainda segundo os autores, o sistema considera diversos fatores quando seleciona tags para sugerir, incluindo a freqüência de ocorrência da tag nas postagens anteriores. O processo inclui, ainda, o estágio que os autores chamaram de compressão das tags, que é composto de duas fases (normalização de tags e validação da compressão). Na fase de normalização, além da retirada de espaços em branco e pontuação, é aplicada a técnica de *stemming*, com o algoritmo de Porter (1980), para retirada de prefixos e sufixos. As tags que contêm mais de uma palavra são "tokenizadas" e colocadas em ordem alfabética. Isso resolve, segundo os autores, variações como "*news and politics*" e "*politics and news*", onde ambos são convertidos para "*and new polit*". Os autores afirmam que essa fase (compressão das tags) foi responsável por uma redução de mais de 18% no número de tags únicas.

A segunda fase da compressão das tags, a avaliação da compressão ou validação serve para confirmar os grupos da fase de normalização e garantir que o sistema não agrupou tags com significados diferentes sob a mesma raiz normalizada.

Os autores advertem que a normalização morfológica é uma técnica relativamente agressiva onde termos que compartilham a mesma raiz, mas possuem significados diferentes podem ser agrupados. Para explicitar esse problema, os autores citam os termos, em inglês, "*production*", "*product*" e "*producers*" que compartilham a mesma raiz "*product*", mas possuem significados diferentes.

Para resolver o problema, os autores investiram na descoberta, para cada tag, de tags relacionadas e definindo aquelas de maior ocorrência como o centróide de cada tag. A partir da comparação dos centróides das tags, as tags são confirmadas ou retiradas de um determinado grupo. Para ilustrar, os autores apresentam as fig. 2.15 e 2.16 onde os termos "*apple*" e "*apples*" não foram enquadrados no mesmo grupo, pois "*apple*" refere-se a um fabricante de tecnologia e "*apples*" se refere à fruta "maçã".

tag	tag relacionada	contagem
apple	Mac	333
apple	Technology	240
apple	iPod	217
apple	Software	190
apple	Microsoft	143
apple	iTunes	135
apples	Fruit	60
apples	Apple	50
apples	Recipes	33
apples	Food	31
apples	Cooking	26
apples	Oranges	20

Fig. 2.13 – Diferença entre *apple* e *apples* - Sood et al.(2007).

Por outro lado, com os termos "*dog*" e "*dogs*" foram considerados participantes de um mesmo grupo, pela etapa de validação, conforme indicado na fig. 2.14

tag	tag relacionada	contagem
dogs	Pets	364
dogs	Dog	108
dogs	Puppies	100
dogs	Cats	82
dogs	Puppy	74
dogs	Dog Training	71
dog	Dogs	108
dog	Pets	93
dog	Puppy	83
dog	Puppies	76
dog	Dog Training	72
dog	Dog Clothes	69

Fig. 2.14 – Diferença entre *dogs* e *dog* - Sood et al.(2007).

Para avaliar a efetividade do sistema TagAssist (Sood et al,2007), foram usados dados de seis dias de 2006 provenientes do site Technorati. Os autores desenvolveram uma versão do sistema sem compressão de tags e avaliação da compressão. Essa versão foi considerada como baseline. As avaliações foram realizadas por avaliadores humanos que aferiram também as tags originais. Os resultados podem ser visualizados na fig. 2.15.

conjunto	precisão
Original Tags	48.85%
TagAssist	42.10%
Baseline	30.05%

Fig. 2.15 – Resultados - Sood et al.(2007).

No total foram coletadas e analisadas 225 respostas de dez avaliadores diferentes. Segundo os autores, o sistema TagAssist apresentou um desempenho significativamente melhor em comparação ao baseline e só foi pior avaliado quando comparado ao desempenho das tags originalmente cadastradas. Segundo os autores, os resultados mostram que o sistema provê relevantes sugestões de tags para novas postagens. O algoritmo proposto, implementado e testado, apresenta uma boa precisão sem perder recuperação e pode proporcionar muitos benefícios às comunidades de blogs. Em primeiro lugar, o sistema proporciona que o registro das tags deixe de ser um processo de geração da informação por parte do usuário para ser um processo de escolha, o que, segundo os autores, reduz significativamente a carga cognitiva sobre os usuários, ou seja, torna o processo mais fácil. Em segundo lugar, conforme os autores, fornecer escolhas de tags baseadas na blogosfera[20] real pode acelerar a convergência de um vocabulário mais consistente e útil para busca e navegação.

[20] Blogosfera é o termo coletivo que compreende todos os weblogs (ou blogs) como uma comunidade ou rede social.(WIKIPÉDIA, 2008)

Specia et al.(2007), apresentam uma abordagem que combina diversas fontes de informação para descobrir conteúdo semântico relacionado as tags e utilizar para:

Expansão da consulta e desambiguação;

Visualização dos *clusters* de tags relacionadas;

Sugestão de tags.

É interessante reforçar que a abordagem utiliza as informações capturadas de Folksonomias como FlickR e del.icio.us de onde são construídos agrupamentos de tags e posteriormente, para confirmar se cada par de tag está realmente relacionado, o sistema utiliza ontologias disponíveis na web, o Google e a Wikipédia. No artigo, os autores demonstram possibilidades de extração de informações dos referidos sistemas, como por exemplo, submeter a tag “nyc” a Wikipédia (http://en.wikipedia.org/wiki/nyc) e obter o resultado (título da página) que será a palavra “New York City”, ou ainda, submeter a busca da tag “sanfrancisco” ao Google (http://www.google.com.br/search?q=sanfrancisco) e obter o resultado (sugestão do Google), “Você quis dizer: San Francisco”.

Nas conclusões, os autores (Specia et al., 2007) afirmam que a abordagem é viável e apresenta resultados promissores.

3. SisRecAC - Sistema de Recuperação de Artigos Científicos

Neste capítulo serão descritos o objetivo, a arquitetura básica do SisRecAC - Sistema de Recuperação de Artigos Científicos, seu funcionamento e principais funções.

3.1 Objetivo

O objetivo do presente trabalho é definir, implementar e avaliar métodos de extração de palavras-chave de documentos textuais para gerar recomendações de artigos científicos. Os métodos de extração de palavras-chave serão comparados a métodos que utilizam os metadados informados pelos usuários, como o título do documento e as tags que foram cadastradas pelos mesmos para identificar o documento. A ferramenta, que pode ser classificada como uma meta-busca, utiliza como fonte para Recuperação o mecanismo de busca chamado Google Acadêmico[21].

.

[21] http://academico.google.com.br

3.2 Arquitetura do Sistema

O sistema desenvolvido possui uma arquitetura típica de sistemas Web, rodando em linux RedHat versão 9 e foi desenvolvido basicamente com a linguagem PHP 5 e o banco de dados PostgreSQL 8.1. Este armazena os dados dos usuários, *stopwords*, dados dos documentos, recomendações e categorias. Para realizar a extração das palavras-chave, foi necessário converter os documentos em formato pdf, enviados pelo usuário, para o formato texto. Essa conversão é realizada pelo software de conversão denominado xpdf, via chamadas ao sistema operacional com o comando "*system*" do PHP.

A figura 3.1 apresenta a estrutura básica do SisRecAC, onde o usuário pode enviar documentos e receber recomendações. Além disso, o usuário pode avaliar as recomendações recebidas. Na figura 3.1 é destacado também que o sistema envia palavras-chave para o Google Acadêmico e recebe como resposta os links para os artigos científicos. Todas as informações são armazenadas na base de dados do sistema.

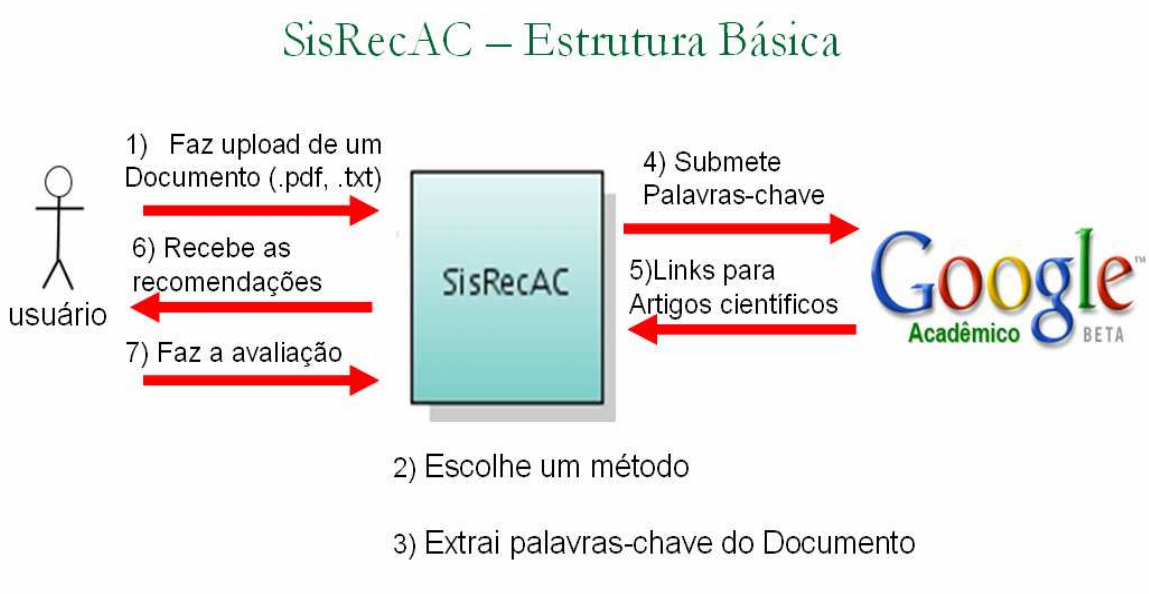

Fig. 3.1 – Estrutura básica do sistema.

3.3 Funcionamento na visão do usuário

Para utilizar o SisRecAC o usuário deve realizar o cadastro no sistema, pois somente usuários autenticados podem utilizar a ferramenta. Após a autenticação no sistema, este pode realizar o envio de documentos, denominados de documento exemplo.

A fig. 3.2 apresenta a tela de abertura do sistema, com o menu na parte superior.

Fig. 3.2 – SisRecAC – Tela de abertura.

3.3.1 Envio de documentos (*upload*), Nuvem de Tags e Listagem

O item Documentos, possui duas funcionalidades principais, permitir o envio de documentos para o sistema e visualizar os documentos já enviados. Ao clicar em Documentos – *Upload*, o usuário terá acesso à tela apresentada na fig. 3.3.

Para realizar o *upload*, o usuário deve informar o título do documento, as tags, algum comentário opcional, o documento propriamente dito e a linguagem para a seleção das *stopwords*, atualmente disponível apenas em inglês e português. Após o preenchimento das informações o usuário deve clicar em “Gravar Arquivo”. Atualmente só é possível enviar arquivos no formato desenvolvido pela empresa Adobe, conhecido por pdf ou então no formato texto (.txt).

Um recurso importante ainda não implementado é a possibilidade de gravação de um texto a partir da digitação direta ou copiar/colar (*cut/past*), a partir de uma fonte qualquer.

Recurso semelhante é oferecido pelo sistema eTBLAST, que foi estudado e descrito na seção 2.1 deste trabalho.

Após a realização do *upload*, o usuário pode visualizar os documentos que foram gravados no sistema e conseqüentemente as recomendações baseadas nesse documento.

Fig. 3.3 – SisRecAC – Tela de Upload.

A fig. 3.4 apresenta a tela do sistema que relaciona os documentos que foram gravados pelo usuário. Na parte superior, logo abaixo do menu, uma nuvem de tags (*tagcloud*) apresenta as tags que foram utilizadas para identificar os documentos gravados pelo usuário no sistema.

A nuvem de tags, no caso do SisRecAC, tem três funções básicas, (1) apresentar, em um formato visualmente agradável, as tags que estão sendo utilizadas e destacando as mais utilizadas (podemos destacar os termos “php”, “recomendação”, “rails”, por exemplo.) (2) servir de filtro para a listagem que está sendo apresentada, bastando que o usuário clique em uma das tags que aparece na nuvem, conforme está ilustrado na fig. 3.5, onde foi utilizado, para a filtragem, um clique no termo “ontology”. No final da listagem é apresentado o total de arquivos que foram identificados pela tag escolhida; e, (3) apresentar o número de vezes que a tag foi utilizada e as tags relacionadas. A fig. 3.6 destaca a apresentação das tags relacionadas, onde são apresentados os termos “web”,”semantic”,”tagging”,”metadata”, “folksonomy”.

A técnica para descobrir as relações entre as tags está baseada na extração de informações de sistemas Web que utilizam Folksonomia e será descrita na seção 3.6.

Fig. 3.4 – SisRecAC – Tela com a relação de arquivos gravados.

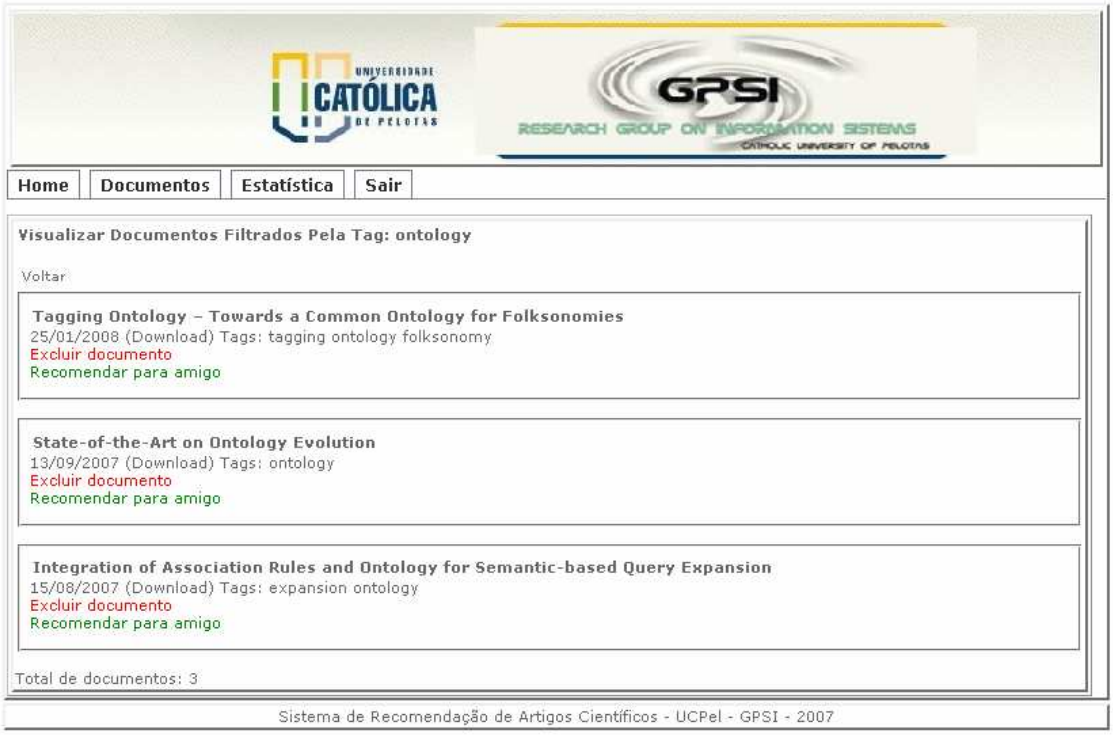

Fig. 3.5 – SisRecAC – Listagem de arquivos identificados com a tag “ontology”.

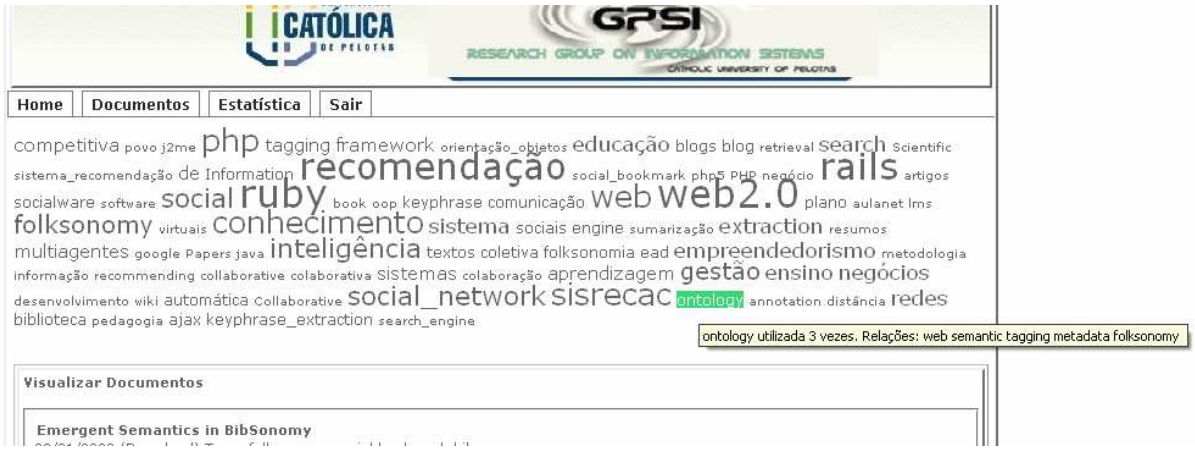

Fig. 3.6 – SisRecAC – Nuvem de tags – Destaque para “tags relacionadas”.

3.3.2 Visualização dos Artigos Recomendados

Para utilizar as recomendações do sistema basta acessar a listagem dos documentos enviados (Fig. 3.4) e clicar no título do documento que foi enviado, por exemplo, em “*Emergent Semantics in BibSonomy*”. A fig. 3.7 apresenta a tela que será disponibilizada ao usuário, ou seja, as recomendações do sistema tendo como base o documento enviado. São apresentadas três recomendações e o usuário é convidado a avaliar cada delas, respondendo

se o artigo recomendado, em relação ao documento exemplo, é: “Totalmente relevante”, “Parcialmente relevante” ou “Irrelevante”.

Na parte inferior da tela (fig. 3.8), o usuário pode clicar em “GERAR NOVA LISTA” para que o sistema apresente mais três recomendações baseadas em um novo método, sorteado pelo sistema. O método que é utilizado para extrair informações do documento exemplo é aleatório e não conhecido pelo usuário.

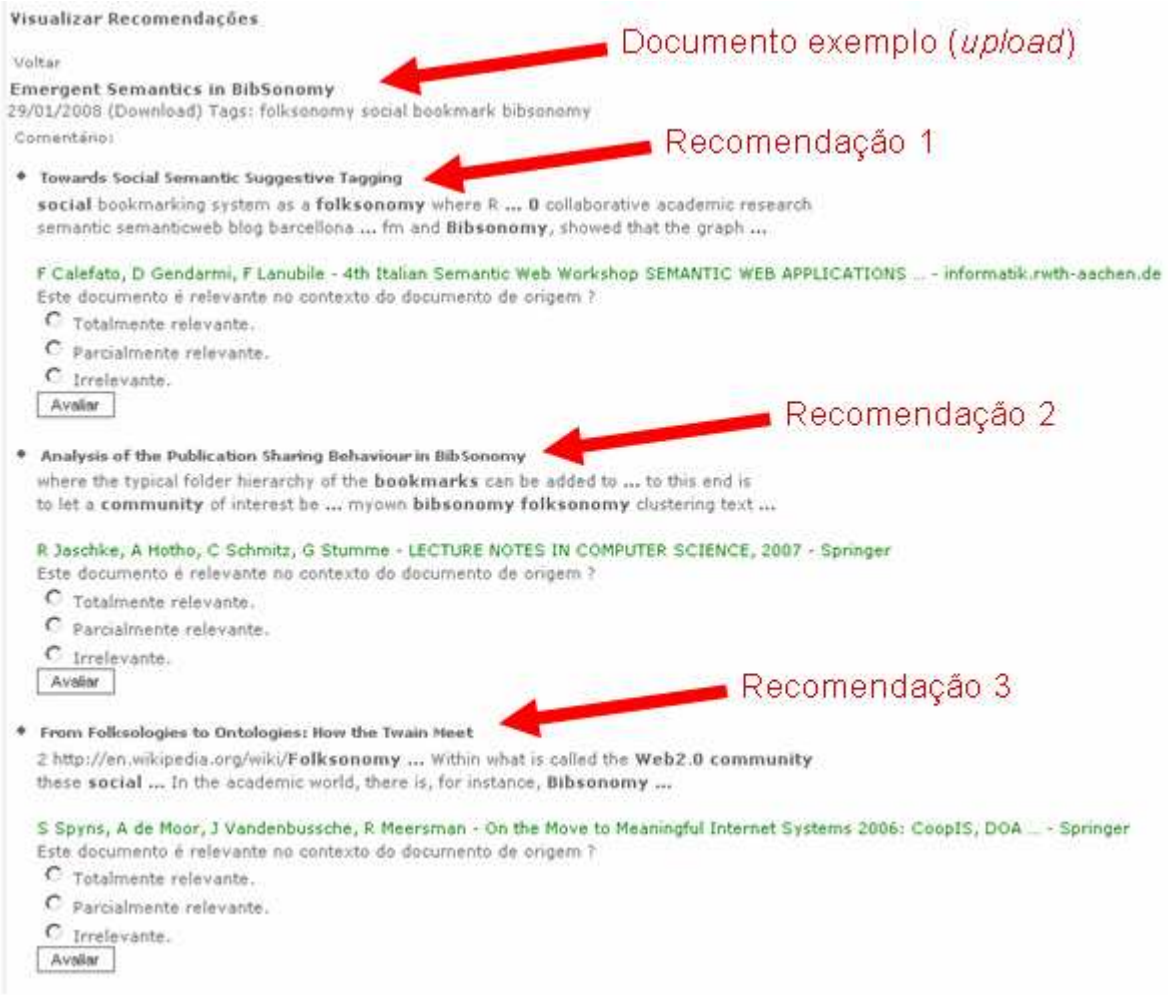

Fig. 3.7 – SisRecAC – Recomendações.

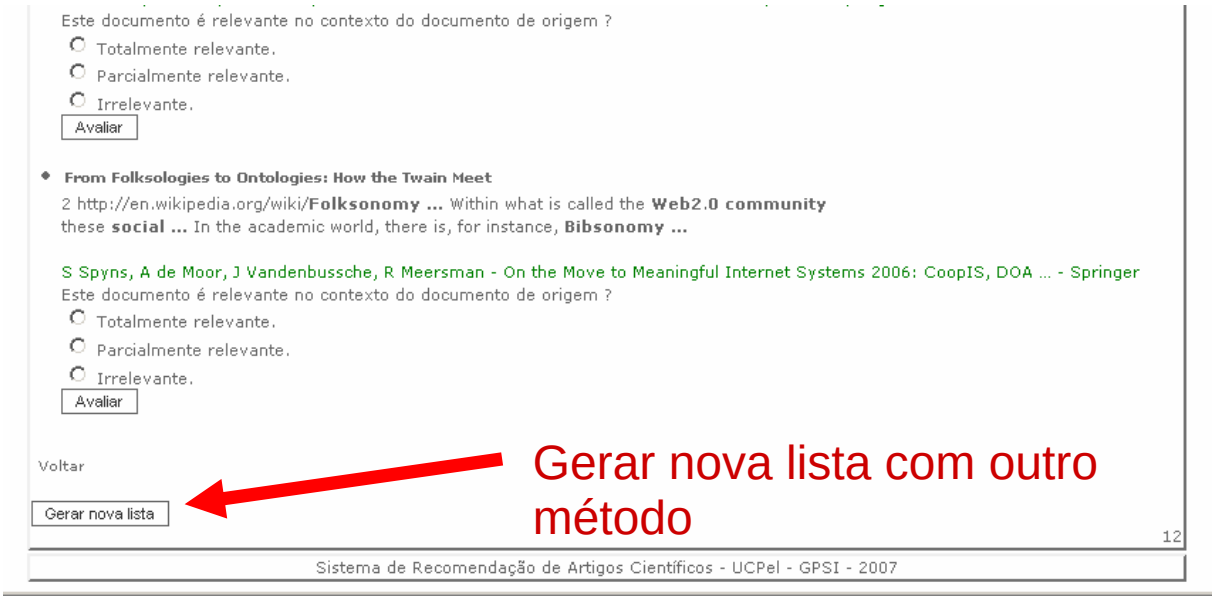

Fig. 3.8 – SisRecAC – Recomendações – Gerar nova listagem.

3.3.4 Estatísticas

As estatísticas do sistema podem ser acessadas em Estatísticas – Precisão dos Métodos. Estão disponíveis alguns dados estatísticos que serão melhor descritos na seção 3.9.

3.4 Principais Funções Internas

A seguir é apresentada uma visão interna do sistema, através da descrição de algumas funções ou programas que foram desenvolvidos. A figura 3.9 apresenta a seqüência de processos entre o envio (*upload*) de um documento exemplo e a avaliação e gravação da avaliação dos links que foram recomendados baseado no documento enviado.

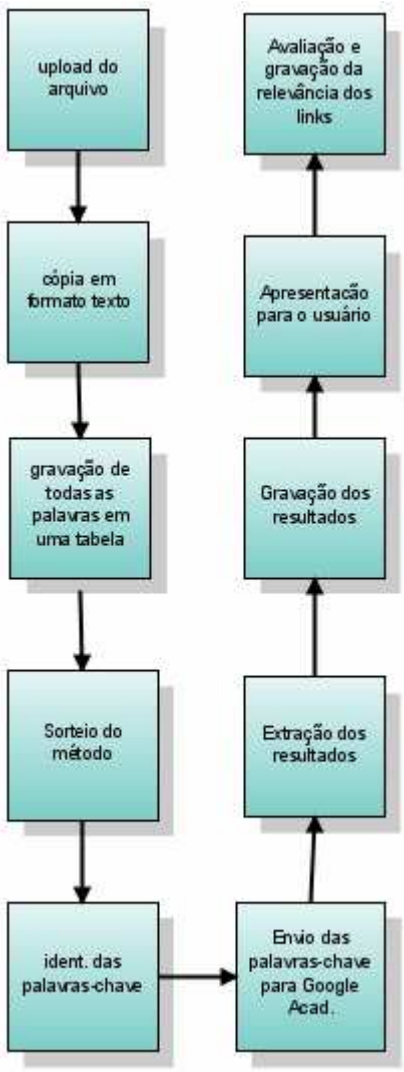

Fig. 3.9 – Seqüência de funções/atividades realizadas pelo sistema.

3.4.1 Envio do Arquivo (Upload)

Trata-se do processo inicial, cabendo ao usuário submeter o documento exemplo (pdf ou txt) ao sistema (Fig. 3.3).

3.4.2 Cópia e Conversão do Documento de PDF para Texto

A partir do momento que o usuário faz o envio (*upload*) do documento exemplo para o servidor, o arquivo original é gravado em um determinado diretório, e uma cópia em formato texto do arquivo é criada pelo sistema. O SisRecAC trabalha atualmente com dois formatos de arquivos (documento exemplo): o formato texto e o formato pdf. Quando o arquivo enviado

está no formato pdf[24], a cópia deve ser o resultado da conversão do arquivo pdf para texto. Para realizar essa conversão (pdf → texto), é utilizado o software xpdf[25] que é um software livre e atende satisfatoriamente à necessidade do sistema em termos de conversão.

3.4.3 Gravação do Arquivo Texto em uma Tabela

Na seqüência, o arquivo no formato texto tem todas suas palavras gravadas em uma tabela (rec_todaspalavras) do banco de dados, onde cada palavra ocupa uma tupla. Essa operação facilita a identificação posterior das palavras-chave do documento.

3.3.4 Sorteio do Método

A cada envio de um documento ou geração de uma nova lista, um dos vinte métodos é escolhido. A escolha do método não é totalmente aleatória, pois o algoritmo que o escolhe verifica qual destes possui o menor número de avaliações, fazendo com que todos os métodos tenham um número semelhante de avaliações. A relação e a descrição detalhada de cada método podem ser consultadas na seção 3.5.3.

3.4.5 Identificação das Palavras-chave

Para extrair as palavras-chave do documento se utiliza o método da freqüência simples, ou seja, um comando SQL retorna em ordem de maior freqüência a quantidade de vezes que cada palavra aparece no documento. Esse grupo de palavras é chamado de palavras-chave (*keywords*) do documento e serão submetidas ao Google Acadêmico. Durante o levantamento da freqüência, o algoritmo acessa uma *stoplist* para que a palavra tenha a sua freqüência desconsiderada caso seja identificada com uma *stopword* que são artigos, pronomes, preposições, advérbios e outros, que não contribuem com a caracterização do texto caso sejam incluídas como palavras-chave.

Para fazer a identificação das palavras que mais se repetem, o algoritmo acessa a tabela rec_todaspalavras, onde estão armazenadas todas as palavras de todos os documentos (documentos base) que foram enviados (*upload*) para o servidor. As palavras-chave

[24] http://www.adobe.com/br/products/acrobat/adobepdf.html

[25] http://www.foolabs.com/xpdf/

identificadas são gravadas em uma tabela chamada rec_dadoslistarecomenda

Na seção 3.5, serão detalhados métodos implementados e avaliados nesse trabalho para extração ou identificação de palavras-chave no texto de entrada.

3.4.6 Envio das Palavras-chave e Extração das Informações do Resultado

Um outro algoritmo recebe as palavras-chave identificadas, submete ao Google Acadêmico e extrai as informações (links, títulos, resumos) do resultado da pesquisa. Para que a extração funcionasse foi necessário analisar as tags geradas pelo resultado em html da pesquisa, para que fosse possível a identificação exata do link, título e resumo. Na Figura 3.10 é possível visualizar um exemplo de html gerado pelo Google Acadêmico.

```
<html><head><meta HTTP-EQUIV="content-type" CONTENT="text/html; charset=UTF-8"><title>keyword extraction - Google Ac
body,td,div,.p,a{font-family:arial,sans-serif}
div,td{color:#000}
.f,.fl:link{color:#6f6f6f}
a:link,.w,a.w:link,.w a:link{color:#00c}
a:visited,.fl:visited{color:#551a8b}
a:active,.fl:active{color:#f00}
.t a:link,.t a:active,.t a:visited,.t{color:#000}
.t{background-color:#dcf6db}
.k{background-color:#008000}
.j{width:34em}
.h{color:#008000;font-size:14px}
.i,.i:link{color:#a90a08}
.a,.a:link{color:#008000}
.z{display:none}
div.n {margin-top:1ex}
.n a{font-size:10pt;color:#000}
.n .i{font-size:10pt; font-weight:bold}
.q a:visited,.q a:link,.q a:active,.q {color:#00c}
.b{font-size:12pt;color:#00c;font-weight:bold}
.ch{cursor:pointer;cursor:hand}
.fl:link{color:#7777cc}
//-->
</style>
</head><body bgcolor="#ffffff"onLoad="document.gs.reset()" topmargin=2 marginheight=2><table border=0 cellpadding=0
there is not many successful works on Chinese <b>keyword</b> <b>extraction</b>. <b>...</b>
<br><font color=#7777CC><a href="/scholar?hl=pt-BR&lr=&cites=18085301927823163147"><font color=#7777CC>Citado por 84
documents, Proceedings of the 25th annual international ACM SIGIR conference on <b>...</b>
<br><font color=#7777CC><a href="/scholar?hl=pt-BR&lr=&cites=8479442330366741056"><font color=#7777CC>Citado por 143
about a page is obtained by means of <b>keyword</b> <b>extraction</b>. <b>...</b>
<br><font color=#7777CC><a href="/scholar?hl=pt-BR&lr=&cites=5156104965904185447"><font color=#7777CC>Citado por 14<
Anette Hulth Department of Computer and Systems Sciences <b>...</b>
<br><font color=#7777CC><a href="/scholar?hl=pt-BR&lr=&cites=875183331021981334"><font color=#7777CC>Citado por 13</
Information <b>...</b> Keywords: <b>keyword</b> <b>extraction</b>, co-occurrence, χ 2 -measure <b>...</b>
<br><font color=#7777CC><a href="/scholar?hl=pt-BR&lr=&cites=9056459318442080509"><font color=#7777CC>Citado por 12<
```

Fig. 3.10- Resultado em html de uma pesquisa realizada no Google Acadêmico.

Um grande problema da abordagem apresentada é a instabilidade que pode ser causada ao sistema caso o código fonte gerado pelo resultado da pesquisa, que é em html, seja alterado

pelos desenvolvedores do Google Acadêmico. Neste caso seria necessária uma nova análise do código fonte para localização das marcas para extração das informações e reprogramação do SisRecAC.

3.4.7 Gravação dos Dados Localizados

Após a extração dos dados dos links localizados pelo Google Acadêmico, ou seja, a ordem do link, o endereço do link, o título, o resumo (*snippet*) e a origem, um outro algoritmo do sistema recebe esses dados e os grava em uma tabela (rec_dadoslinks) no banco de dados para que posteriormente possa ser apresentado e avaliado pelo usuário.

Fig. 3.11 - Dados que são extraídos e gravados na tabela rec_dadoslink.

3.4.8 Apresentação do Resultado e Avaliação da relevância de cada link

Após a extração das informações, uma tela apresenta essas informações ao usuário (Fig. 3.12) para que este avalie a relevância de cada link.

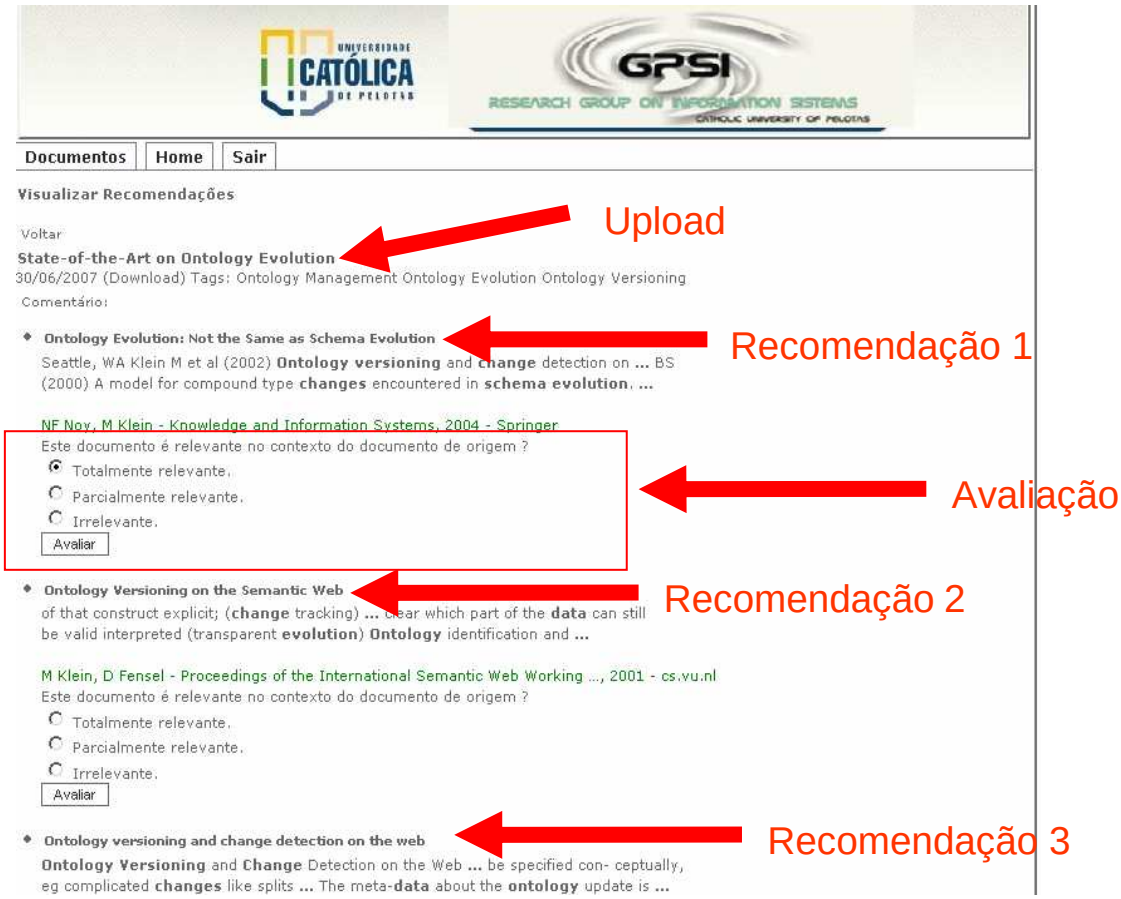

Fig. 3.12 – Apresentação do resultado e avaliação subjetiva de relevância.

3.4.9 Modelo ER da Base de Dados

Os dados estão armazenados em um SGBD PostgreSQL 8.1.4 e a A Figura 3.13 apresenta o modelo ER do sistema.

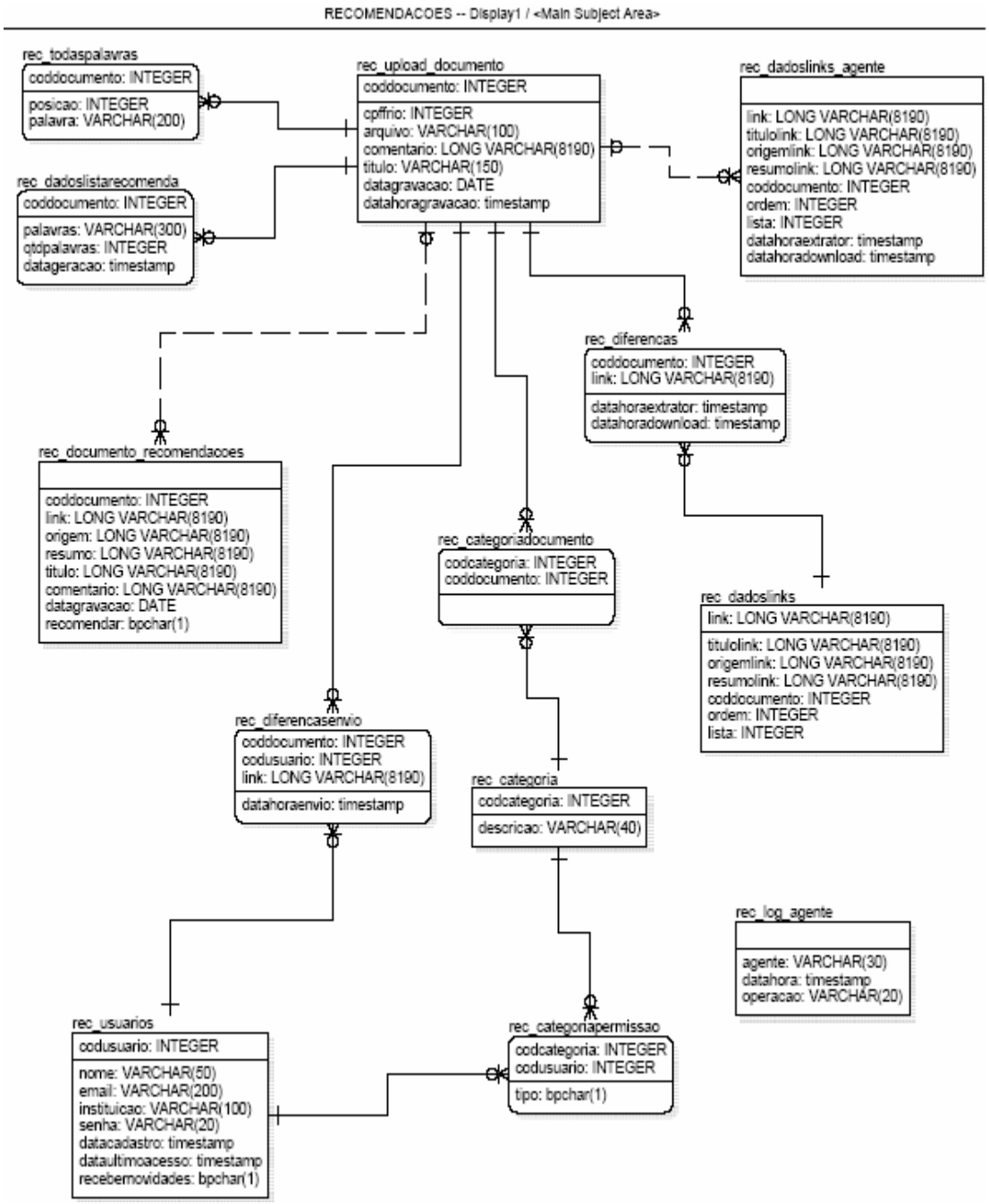

Fig. 3.13 – Modelo ER do SisRecAC

3.5 Métodos Avaliados no SisRecAC

Neste tópico serão descritos os métodos do SisRecAC que foram implementados e avaliados.

3.5.1 Métodos de Descoberta de Expressões

O algoritmo desenvolvido, que implementa o método de descoberta de expressões em documentos, analisa a freqüência com que duas ou três palavras aparecem juntas no texto. Os métodos, que submetem ao Google Acadêmico expressões extraídas do documento, enviam apenas expressões, ou seja, duas ou três palavras entre aspas (").

É importante observar que no caso das expressões, as *stopwords* são consideradas apenas quando estão entre dois termos, como em "banco de dados" e não são consideradas como válidas quando no início ou fim de uma expressão, como por exemplo, em "a descoberta de" ou "a inteligência artificial". Isso porque no primeiro caso, não teríamos uma expressão, mas sim apenas o termo "descoberta". Já no segundo, teremos a expressão "inteligência artificial".

Fazem parte desse tipo de abordagem os métodos Expr1, Expr2 e Expr3 que são melhor descritos na seção 3.5.3.

3.5.2 Métodos de Freqüência de Termos Simples

Nós métodos de freqüência de termos simples, o algoritmo identifica os termos simples (não expressões) que mais se repetem. Nesse caso, são excluídas as *stopwords*, que podem ser em português ou inglês. O próprio usuário, no momento do upload, informa o idioma do documento e a informação é utilizada para selecionar a *stoplist*. Posteriormente, os termos de maior freqüência são concatenados e enviados ao Google Acadêmico. Os métodos deste tipo variam a quantidade de termos que devem ser concatenados. Em uma fase preliminar desse trabalho, foram realizadas avaliações informais. Os resultados apontaram para um melhor desempenho na faixa entre quatro e nove termos. Portanto para esse tipo método foram realizados testes com o envio de 4, 5, 6, 7, 8 e 9 termos.

A tabela 3.1 apresenta uma relação hipotética de palavras-chave extraídas de um documento, na qual o termo “banco” aparece quinze vezes no texto, juntamente com o termo “dados e assim sucessivamente até o termo “processos” que aparece três vezes. Caso o método de extração escolhido fosse de freqüência simples para quatro termos, as palavras-

chave escolhidas seriam: banco, dados, gerenciamento e projeto, e a url a ser submetida seria (simplificadamente): http://scholar.google.com.br/scholar?q=banco+dados+gerenciamento+projeto.

Utilizam puramente a abordagem descrita nessa seção os métodos PC4, PC5, PC6, PC7, PC8, PC9 e também, mas de forma combinada com outras técnicas, os métodos PC2Exp2, PC3Exp3, PC4Exp4, PC2ORExp2, PC3ORExp3 e PC4ORExp4.

Tabela 3.1 – Relação hipotética de palavras-chave.

PALAVRA	**FREQÜÊNCIA**
banco	15
dados	15
gerenciamento	12
projeto	11
modelagem	8
sgbd	7
postgresql	6
relacional	5
sistema	4
processos	3

3.5.3 Relação Geral dos Métodos

A seguir serão descritos os vinte métodos que foram implementados e testados no SisRecAc. Para facilitar a identificação, cada método foi nomeado com uma sigla e um número, que possuem alguma relação com o tipo de método, onde Expr significa expressões, PC significa palavras-chave, Tit significa título e Tag significa tag. Os números identificam a quantidade de palavras-chave ou termos expandidos. A tabela 3.2 apresenta uma descrição resumida de todos os métodos.

Método Expr1 - O algoritmo extrai a expressão de maior freqüência no documento. A expressão pode ter duas ou três palavras. Por exemplo: "banco de dados".

Método Expr2 - Igual ao método anterior, porém são identificadas duas expressões de maior freqüência. Por exemplo: "banco de dados"+"inteligência artificial" .

Método Expr3 - Idem dois primeiros, porém com três expressões. Por exemplo: "banco de dados"+"inteligência artificial"+"gerência de projetos".

Método PC4 - O método PC4 não identifica expressões, e sim as quatro palavras de maior freqüência no documento. Por exemplo: banco+dados+inteligência+artificial.

Métodos PC5, PC6, PC7, PC8, PC9 - Estes métodos diferem do método anterior apenas pelo número de palavras que são extraídas (5, 6, 7, 8 e 9 palavras respectivamente).

Método Tit - O método Tit submete o título ao mecanismo de busca (retiradas as *stopwords*) que foi cadastrado pelo usuário do SisRecAC. Apesar de o título fazer parte do documento, optou-se por utilizar a informação que é cadastrada pelo usuário a fim de simplificar o processo, já que desenvolver um algoritmo de extração seria um processo complexo tendo em vista a natureza diversa dos documentos exemplos.

Método Tag - O método Tag submete ao mecanismo de busca as tags que foram cadastradas pelo usuário do SisRecAC para identificar o documento. A análise dos resultados desse método, possibilitou uma comparação com os resultados apresentados por Brooks e Montanez (2006) que analisam o desempenho da geração automática de tags em comparação

às tags que foram definidas pelos usuários.

Método TagExp1 –Utiliza a expansão semântica para acrescentar novas tags as tags que foram informadas pelo usuário. Para cada tag associada ao documento, uma nova tag é acrescentada. A seção 3.6 descreve como as novas tags são descobertas para uso na expansão semântica. Estas são acrescentadas a url que é submetida ao Google Acadêmico com o operador AND, que significa que TODOS os termos enviados devem aparecer no documento recomendado. Por exemplo, se o usuário identificou o documento com a tag "*framework*", o sistema deve incluir o termo "web", ficando a url que faz a busca da seguinte forma (simplificadamente, pois alguns parâmetros foram omitidos): http://scholar.google.com.br/scholar?q=framework+web, onde o símbolo + significa AND.

Método TagExp2 – Similar ao método anterior. Porém para cada tag são acrescentadas duas novas tags.

Método TagExp3 – Similar ao método anterior. Todavia para cada tag são acrescentadas três novas tags.

Método PC2Exp2 – Utiliza a expansão semântica para acrescentar uma nova tag às palavras-chaves que foram extraídas do texto. O método PC2Exp2 parte das duas palavras-chave mais freqüentes para realizar a expansão, ficando no total quatro termos a serem submetidos a busca no Google Acadêmico. Este é o número mínimo de termos com resultados razoáveis segundo avaliações preliminares realizadas no escopo deste trabalho.

Método PC3Exp3 – Este método parte das três palavras-chave mais freqüentes e realiza a expansão em uma tag para cada palavras-chave, totalizando em seis termos a serem submetidos a busca.

Método PC4Exp4 – Idem aos métodos PC2Exp2 e PC3Exp3, porém partindo com quatro palavras, totalizando oito termos

Método PC2ORExp2 – Todos os métodos anteriores utilizam apenas o operador AND (+) para submeter os termos ao Google Acadêmico, fazendo com que os artigos retornados tenham obrigatoriamente todos os termos submetidos. O operador OR foi testado nos métodos PC2ORExp2, PC3ORExp3 e PC4ORExp4, ou seja, a mesma expansão do método PC2Exp2, porém com o operador OR. Partindo de duas palavras-chave mais

freqüentes, cada uma das palavras é expandida com uma tag. Por exemplo, para o caso de um documento identificado com as tags “*framework*” e “php”, a expansão seria, respectivamente com as tags “web” e “programming”, ficando a url no seguinte formato:

http://scholar.google.com.br/scholar?q=framework+OR+web+AND+php+OR+programming

A submissão dessa url pode recuperar documentos, por exemplo, que possuam apenas os termos “web” e “programming”, que são termos que possivelmente nem apareçam no documento exemplo do usuário, mas estão, de alguma forma, relacionados com as palavras-chave extraídas. Esse fato pode levar o sistema a recomendar documentos inesperados (*Serendipity*), que seria algo positivo na visão de alguns autores (McNee et al.2006; Adomavicius & Tuzhilin, 2005)

Método PC3ORExp3 – Semelhante ao método PC2ORExp2, porém partindo de 3 palavras-chave que são expandidas com o operador OR.

Método PC4ORExp4 – Semelhante aos métodos PC2ORExp2 e PC3ORExp3, porém partindo de quatro palavras-chave.

Dos vinte métodos, quinze utilizam de alguma forma palavras extraídas do documento exemplo. Os demais (5) utilizam apenas informações cadastradas pelo usuário (título ou tags) mais expansão semântica baseada em Folksonomias. Esses métodos (Tit, Tag, TagExp1, TagExp2, TagExp3), que utilizam informações cadastradas pelo usuário são extremamente importantes no contexto do SisRecAC, já que muitos documentos enviados pelos usuários são protegidos e não podem ter as palavras-chave extraídas. Portanto, mesmo que os resultados indiquem que os melhores métodos são aqueles cujas palavras-chave são extraídas automaticamente pelo sistema, um ou mais dos outros métodos serão utilizados quando o documento exemplo estiver bloqueado.

Tabela 3.2 – Resumo da descrição de cada método.

MÉTODO	DESCRIÇÃO (RESUMO)
Expr1	A expressão de maior freqüência
Expr2	As Duas expressões de maior freqüência
Expr3	As Três expressões de maior freqüência
PC4	As Quatro palavras de maior freqüência
PC5	As Cinco palavras de maior freqüência
PC6	As Seis palavras de maior freqüência
PC7	As Sete palavras de maior freqüência
PC8	As Oito palavras de maior freqüência
PC9	As Nove palavras
Tit	Palavras do Título (sem stopwords)
Tag	Tags cadastradas pelo usuário
TagExp1	Tags do usuário com expansão de uma palavra para cada tag
TagExp2	Tags do usuário com expansão de duas palavras para cada tag
TagExp3	Tags do usuário com expansão de três palavras para cada tag
PC2Exp2	Duas palavras de maior freqüência com expansão de uma palavra para ca
PC3Exp3	Três palavras de maior freqüência com expansão de uma palavra para ca
PC4Exp4	Quatro palavras de maior freqüência com expansão de uma palavra para
PC2ORExp2	Duas palavras de maior freqüência com uma expansão para cada, porém o operador OR entre palavra e expansão
PC3ORExp3	Três palavras de maior freqüência com uma expansão para cada (OR)
PC4ORExp4	Quatro palavras de maior freqüência com uma expansão para cada (OR)

3.6 Descoberta de tags relacionadas a partir de Folksonomias

No SisRecAC, os textos são identificados pelos usuários por meio de uma ou mais tags que são livremente informadas para cada documento.

As tags cadastradas são utilizadas na montagem de uma nuvem de tags (*tagcloud)* que facilita a busca de artigos e apresentam de forma explícita as tags mais utilizadas. Na fig. 3.4 pode ser observado que a tag “Recomendação” foi utilizada mais vezes que a tag “virtuais”. Além disso, ao passar com o *mouse* por uma tag, o sistema informa exatamente o número de vezes que aquela tag foi utilizada pelo usuário e também as relações, ou seja, tags que estão relacionadas a àquela tag. No caso da fig. 3.6, o *mouse* está sobre a tag “*ontology*” e o sistema informa que as tags relacionadas são “*web*”, ”*semantic*”, ”*tagging*”, ”*metadata*” e ”*folksonomy*”.

Tais relações são descobertas por um algoritmo que acessa diversos sistemas que utilizam Folksonomias, captura as relações e grava em uma tabela do SisRecAC. As relações são utilizadas na visualização e também na expansão semântica das tags informadas pelos usuários. Estas são extraídas dos sites que apresentam a informação “tags relacionadas” (*related tags*).

A fig. 3.13 ilustra um exemplo de um site de *social bookmark*, o Bibsonomy, o qual utiliza Folksonomia, é uma das fontes utilizada pelo algoritmo para descobrir tags relacionadas. No exemplo da figura (fig. 3.13), foi feita uma busca pela tag “framework”. O sistema retornou os diversos links que foram identificados por essa tag, mas a informação que é capturada pelo algoritmo do SisRecAC está destacada a esquerda, são as “Related Tags”. O algoritmo extrai informações de diversos sistemas semelhantes ao Bibsonomy, como por exemplo o site del.icio.us (Fig. 3.15).

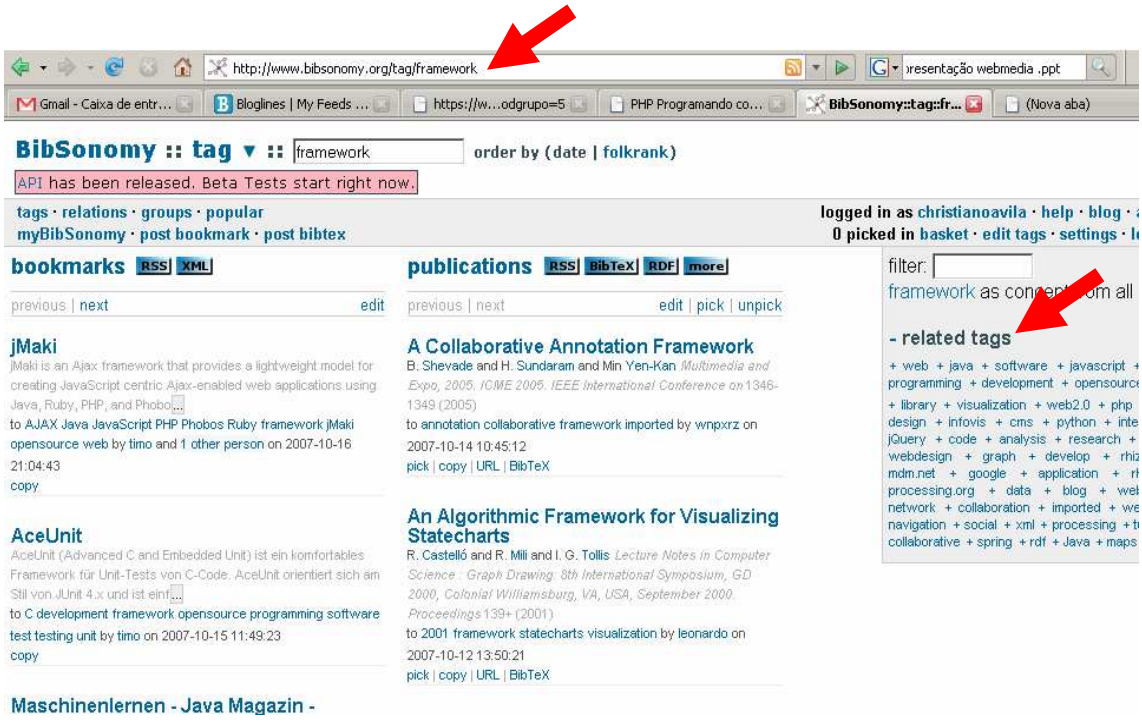

Fig. 3.14 – Bibsonomy – Tags relacionadas (*related tags*).

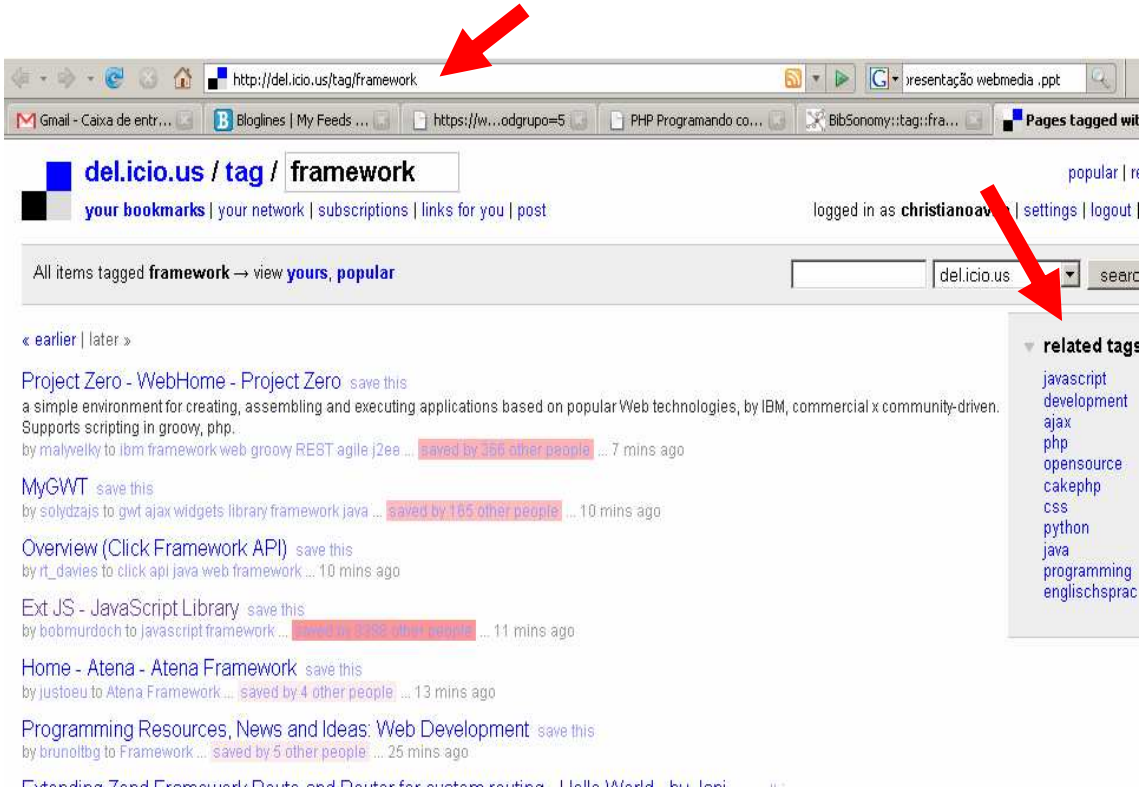

Fig. 3.15 – Del.icio.us – Tags relacionadas (*related tags*).

A fig. 3.16 apresenta uma demonstração onde pode ser observado que a tag “framework” é submetida ao algoritmo que extrai as tags relacionadas de diversos sistemas. Estão relacionados os sistemas e logo abaixo as tags arroladas que foram extraídas. Na parte inferior da figura são apresentadas duas *tagclouds*, uma em ordem alfabética e outra em ordem decrescente do número de vezes que a tag aparece relacionada a “*framework*”. Pode ser observado que, segundo a Folksonomia, *framework* tem forte relação com “*web*”, “*development*”, “*programming*”e “*ajax*”.

Fig. 3.16 – Algoritmo de Descoberta de Relações.

Para extrair as informações de cada site, o algoritmo consulta uma tabela (rec_tagfontes) na qual estão cadastradas as marcas que devem ser consideradas para início e fim da extração de cada site.

```
sisrecac=# select linkhome,stringinicio,stringfim from rec_tagfontes where ativo='S';
          linkhome          |                stringinicio                |          stringfim
----------------------------+--------------------------------------------+----------------------------
 http://www.bibsonomy.org   | <span class="sidebar_h">related tags</span> | </ul>
 http://www.stumbleupon.com | <div class="pdgTop tagcloud">               | </ul>
 http://bluedot.us/         | +</a>                                       | </a></li></ul></div></div>
 http://yahoo.com/          | <Result>                                    | </ResultSet>
 http://del.icio.us/        | <strong>Related tags:</strong>              | </a></p>
 http://www.tagpatterns.com | <h3>Forward Relationships</h3><ul><li>      | </ul>
(6 rows)
```

Fig. 3.17 –Fontes de informação para tags relacionadas.

Posteriormente, o algoritmo armazena as informações extraídas (tagorigem, site, tagrelacionadas, etc) na tabela rec_tagdados. A fig. 3.18 apresenta um comando *select* que mostra algumas tuplas cuja tag origem é *framework*.

```
sisrecac=# select * from rec_tagdados where tagorigem ='framework';
   id    | tagorigem |  nomesite   |      tag       |          datahora
---------+-----------+-------------+----------------+----------------------------
 7512520 | framework | bibsonomy   | java           | 2008-02-03 21:13:19.639407
 7512521 | framework | bibsonomy   | web            | 2008-02-03 21:13:19.679524
 7512522 | framework | bibsonomy   | software       | 2008-02-03 21:13:19.690486
 7512523 | framework | bibsonomy   | programming    | 2008-02-03 21:13:19.69645
 7512524 | framework | bibsonomy   | javascript     | 2008-02-03 21:13:19.70244
 7512525 | framework | bibsonomy   | ajax           | 2008-02-03 21:13:19.708427
 7512526 | framework | bibsonomy   | opensource     | 2008-02-03 21:13:19.714567
 7512527 | framework | bibsonomy   | library        | 2008-02-03 21:13:19.720555
 7512528 | framework | bibsonomy   | development    | 2008-02-03 21:13:19.72655
 7512529 | framework | bibsonomy   | visualization  | 2008-02-03 21:13:19.732544
 7512530 | framework | bibsonomy   | dev            | 2008-02-03 21:13:19.738528
 7512531 | framework | bibsonomy   | tools          | 2008-02-03 21:13:19.744508
 7512532 | framework | bibsonomy   | php            | 2008-02-03 21:13:19.750505
 7512533 | framework | bibsonomy   | web2.0         | 2008-02-03 21:13:19.756486
 7512534 | framework | bibsonomy   | develop        | 2008-02-03 21:13:19.762485
 7512535 | framework | bibsonomy   | 2.0            | 2008-02-03 21:13:19.768466
 7512536 | framework | bibsonomy   | infovis        | 2008-02-03 21:13:19.774469
 7512537 | framework | bibsonomy   | code           | 2008-02-03 21:13:19.780456
 7512538 | framework | bibsonomy   | design         | 2008-02-03 21:13:19.786451
 7512539 | framework | bibsonomy   | graph          | 2008-02-03 21:13:19.792433
 7512540 | framework | bibsonomy   | cms            | 2008-02-03 21:13:19.798429
 7512541 | framework | bibsonomy   | interface      | 2008-02-03 21:13:19.804411
```

Fig. 3.18 –Armazenamento de tags relacionadas.

Depois de armazenadas as relações, uma visão (*view*) foi criada no banco de dados para facilitar a visualização e utilização no SisRecAC. Tal visão já totaliza e agrupa por tag relacionada. A fig. 3.19 apresenta um comando *select* nessa visão para descobrir as tags mais relacionadas com “*framework*”.

```
sisrecac=# select * from vtags where  tagorigem ='framework';
 count |     tag     | tagorigem
-------+-------------+-----------
     4 | ajax        | framework
     4 | php         | framework
     4 | development | framework
     4 | web         | framework
     4 | programming | framework
     3 | java        | framework
     3 | javascript  | framework
     3 | opensource  | framework
     2 | .net        | framework
     2 | api         | framework
(10 rows)
```

Fig. 3.19 – Visão - Agrupamento por tags relacionadas para tag “framework”.

3.7 Experimentos e Avaliações

Neste capítulo é descrita a metodologia utilizada para realizar os experimentos com os métodos propostos. A avaliação está dividida em duas partes, uma subjetiva. de relevância e outra objetiva, através da aplicação de uma fórmula de similaridade.

3.7.1 Avaliação de Sistemas de Recomendação

Segundo Herlocker et al. (2004), a exatidão dos sistemas de recomendação tem sido alvo de pesquisas desde 1994 e constata ainda que diferentes métricas têm sido utilizadas para avaliar os sistemas de recomendação. No referido estudo, foram analisadas as métricas mais populares e classificadas em três classes: métrica de exatidão da previsão, métrica de exatidão da classificação e métrica de exatidão do ranking.

Segundo (Herlocker et al., 2004) a métrica de exatidão da previsão mede o quanto estão próximas as avaliações previstas pelo sistema com as avaliações verdadeiras do usuário.

Como exemplo, o autor cita o sistema de recomendação chamado MovieLens (Dahlen et al. 1998) que prevê o número de estrelas que um usuário deverá atribuir para cada filme e mostra a previsão para o usuário. A métrica de exatidão da previsão avalia o número de estrelas previstas ao número de estrelas realmente atribuídas pelo usuário a um determinado filme.

Quanto às métricas de exatidão da classificação, o autor explica que esse tipo de avaliação mede a freqüência com que um sistema de recomendação toma decisões corretas ou incorretas sobre se um artigo é bom. Como exemplos de medidas de exatidão da classificação, (Herlocker et al., 2004) cita precisão e abrangência e explica que essas medidas são computadas de uma tabela 2x2 onde os itens são separados em duas classes - relevantes ou não relevantes - e afirma que se a avaliação não for binária, é necessário proceder a conversão para uma escala binária. Cita ainda o sistema MovieLens, que possui uma escala de um (1) a cinco (5), os itens relevantes são aqueles com nota entre quatro e cinco e os não relevantes possuem nota entre um e três. Para precisão e abrangência, o autor afirma ainda que também é necessário separar os itens em um conjunto que foi selecionado/recomendado pelo usuário e em outro aqueles que não o foram.

Dessa forma, é possível calcular a precisão, que é definida como a relação entre a quantidade de itens relevantes selecionados e a quantidade de itens selecionados. Com isso a precisão representa a probabilidade de um item selecionado ser relevante.

$$P = \frac{N_{rs}}{N_s}. \tag{3.1}$$

Fig. 3.20 –Fórmula da precisão (Herlocker et al., 2004).

Na Figura 3.20 é apresentada a fórmula da precisão, que é a relação entre a quantidade de itens relevantes selecionados e a quantidade de itens selecionados.

Já a abrangência Figura 3.21 é definida como a relação entre a quantidade de itens relevantes selecionados e a quantidade de itens relevantes disponíveis. A abrangência representa a probabilidade de um item relevante ser selecionado.

$$R = \frac{N_{rs}}{N_r}. \tag{3.2}$$

Fig. 3.21 - Fórmula da abrangência (Herlocker et al., 2004).

O autor afirma que precisão e abrangência dependem da separação dos itens em relevantes e não relevantes, porém essa separação é subjetiva, variando de usuário para usuário. O autor conclui que essa subjetividade é maior em sistemas de recomendação do que em sistemas tradicionais de recuperação de informação.

A outra métrica apresentada por Herlocker (Herlocker et al., 2004) é a Exatidão do Ranking, que mede a habilidade do sistema de recomendação em apresentar uma listagem na qual os itens estão ordenados de acordo com a preferência do usuário. Diversos estudos foram e estão sendo realizados na área de apresentação de *rankings*.

Em (McNee et al.2006) é exposta a dificuldade de se avaliar sistemas de recomendação, assim como são feitas críticas às métricas que estão sendo utilizadas nessa área. Os autores exemplificam o problema citando o exemplo de um sistema de recomendação de pacotes turísticos, que recomenda apenas os locais que a pessoa já visitou. Mesmo que o sistema recomende os locais perfeitamente, em ordem de preferência e com exatidão será considerado um sistema pobre em termos de recomendação, pois não apresenta novidades a seus usuários. Seguindo nessa linha, o autor cita o termo *Serendipity*, que podemos traduzir no caso como sendo a capacidade de recomendar itens inesperados.

Esse mesmo problema também é destacado em Adomavicius & Tuzhilin (2005), porém os autores utilizam o termo "*Overspecialization*" ou em português "excesso de especialização" e exemplificam com sistemas de recomendação que indicam a seus usuários notícias muito semelhantes de um mesmo evento. Entretanto, segundo os autores, alguns sistemas já fazem não somente a filtragem de itens muito diferentes do perfil do usuário, mas também de itens que são muito similares a algo que o usuário já tenha visto antes a fim de evitar a redundância. Uma das questões está em detectar se uma nova notícia ou documento possui novas e relevantes informações.

Especificamente o tema novidade e redundância em sistemas de recomendação foi foco do trabalho de Zhang et al (2002), onde um conjunto de cinco diferentes medidas de redundância são propostas e avaliadas em experimentos com e sem limiares de redundância. Um documento redundante, na definição dos autores, é um documento relevante cujo conteúdo já foi abordado por outros documentos relevantes que foram previamente recomendados ao usuário. Os autores afirmam que é possível identificar redundância com razoável precisão. Nas avaliações, das cinco medidas de redundância, as de melhores resultados foram conseguidas através da métrica do cosseno de similaridade (Lee, 1999) e uma nova métrica que foi desenvolvida pelos autores, e é baseada em uma combinação métodos de modelagem de linguagem (*language model*).

Em McNee et al.(2006), os autores citam outro trabalho (Ziegler et al.,2005), onde são apresentados métodos para aumentar a satisfação do usuário em relação ao sistema de recomendação, mas em detrimento da precisão média do sistema. Os autores afirmam que as pesquisas em relação a avaliação de sistemas de recomendação, devem ir além da precisão pura e avaliar a experiência do usuário em relação ao sistema para poder obter avanços mais significativos.

3.7.2 Avaliação Subjetiva de Relevância no SisRecAC

A avaliação subjetiva dos métodos será realizada pelos usuários da ferramenta, que irão avaliar cada documento localizado e recomendado. Os usuários não saberão qual o método que foi utilizado pela ferramenta para recomendar o documento. São recomendados apenas três artigos para cada documento exemplo. A precisão foi calculada para o primeiro documento recomendado (@1) e para os 3 primeiros recomendados (@3), seguindo a mesma metodologia utilizada em Ribeiro-Neto et al (2005).

Esse tipo de avaliação subjetiva, contando com a opinião dos usuários, é defendida por alguns autores. Segundo Herlocker,2004, precisão e abrangência dependem da separação dos itens em relevantes e não relevantes, porém essa separação é subjetiva, variando de usuário para usuário. Já Ziegler et al (2005) afirmam que as avaliações de sistemas de recomendação devem ir além das métricas tradicionais e deve-se levar em conta a experiência do usuário em relação ao sistema.

Para avaliar as recomendações, no SisRecAC, o usuário deverá responder as seguintes questões:

a) Este documento é relevante no contexto do documento de origem?

( ) Totalmente Relevante no contexto do documento de origem

( ) Parcialmente Relevante

( ) Irrelevante

Essa pergunta possibilitará uma avaliação da precisão do método, dada pelas seguintes fórmulas:

Percentual de documentos totalmente relevantes

(3.3)

$$\text{PercRelevantes} = \frac{\text{Qtd documentos considerados totalmente relevantes}}{\text{Qtd documentos recomendados}} \text{ X } 100$$

Percentual de documentos parcialmente relevantes

(3.4)

$$\text{PercParcRelevantes} = \frac{\text{Qtd documentos considerados parcialmente relevantes}}{\text{Qtd documentos recomendados}} \text{ X } 100$$

Percentual de documentos irrelevantes

(3.5)

$$\text{PercIrrelevantes} = \frac{\text{Qtd documentos considerados irrelevantes}}{\text{Qtd documentos recomendados}} \text{ X } 100$$

A análise será realizada para cada um dos métodos citados no item 3.5. O usuário não terá a informação do método de busca que foi utilizado para enviar as palavras-chave ao Google Acadêmico.

Para tabular os resultados, serão apresentadas tabelas e gráficos, que serão utilizados na realização de uma análise das precisões médias de todos os métodos.

3.7.3 Avaliação Objetiva (Similaridade) no SisRecAC

O objetivo da avaliação matemática é valorar a similaridade de cada documento recomendado pela ferramenta em relação ao documento exemplo. Esses resultados são interessantes para avaliar se a similaridade é importante (e em que grau de importância) na comparação com a relevância julgada pelo usuário na avaliação subjetiva.

Para a avaliação matemática, utilizou-se uma fórmula de similaridade definida em (Loh 2001). Cconforme o autor, o grau de similaridade entre dois textos é dado pela soma dos graus de igualdade dos termos comuns, dividido pelo número total de termos nos dois documentos.

$$gs(X,Y) = \frac{\sum_{h=1}^{k} gi_h(a,b)}{n} \qquad (3.6)$$

onde:
gs é o grau de similaridade entre os documentos ***X*** e ***Y***;
gi é o grau de igualdade entre os pesos do termo ***h*** (peso ***a*** no documento ***X*** e peso ***b*** no documento ***Y***);
h é um índice para os termos comuns aos dois documentos;
k é o número total de termos comuns aos dois documentos;
n é o número total de termos nos dois documentos (sem contagem repetida).

Fig. 3.22 – Fórmula de Similaridade (Loh 2001).

Para realizar a avaliação matemática foi necessário realizar o *download* e a conversão para formato texto de cada artigo recomendado pela ferramenta. Essa necessidade gera dois problemas: (1) alguns artigos recomendados não estão disponíveis no link indicado pelo Google Acadêmico; e, (2) alguns artigos no formato pdf são protegidos e não podem ser convertidos para o formato texto. Em ambos casos, esses artigos são ignorados e sua similaridade com o documento exemplo não é computada.

Após o *download* e conversão, todas as palavras do artigo são gravadas em uma tabela do banco de dados para que, posteriormente, um algoritmo que implementa a fórmula proposta por Loh (2001) seja executado. Esse algoritmo grava em uma outra tabela o valor da similaridade entre o documento exemplo e o artigo recomendado (fig. 3.23).

```
sisrecac=#  select * from similaridade;
 id_documento | id_recomendacao | posicao | metodo |   similaridade
--------------+-----------------+---------+--------+------------------
           47 |           15659 |       2 |      2 |   0.056571706812
           49 |           15663 |       1 |      5 |   0.052779390388
           49 |           15664 |       2 |      5 |   0.073661556458
           50 |           15666 |       1 |      6 |   0.475256004459
           50 |           15667 |       2 |      6 |   0.234893280569
           50 |           15668 |       3 |      6 |   0.123659834052
           51 |           15671 |       3 |      7 |   0.091435365303
           52 |           15672 |       1 |      8 |  0.0928605024147
           53 |           15675 |       1 |      9 |   0.510421448114
           54 |           15684 |       1 |      3 |   0.502562236262
           55 |           15689 |       1 |      5 |   0.107077386848
           55 |           15690 |       2 |      5 |   0.102306204575
           55 |           15691 |       3 |      5 |   0.132090507778
           56 |           15693 |       2 |      6 | 0.00556073243207
           55 |           15695 |       1 |      7 |    0.50501597024
           55 |           15696 |       2 |      7 |   0.132090507778
           55 |           15697 |       3 |      7 |   0.120041965958
           57 |           15738 |       1 |      1 |  0.0279336484095
           57 |           15739 |       2 |      1 |  0.0153359485608
           57 |           15746 |       1 |      1 |  0.0279336484095
           59 |           15771 |       3 |      5 |  0.0812029522603
           61 |           15779 |       3 |      7 |   0.109507108451
           63 |           15785 |       1 |     10 |  0.0213545805648
           63 |           15786 |       2 |     10 |  0.0102272653631
           62 |           15789 |       1 |     11 |  0.0233772312184
           62 |           15791 |       3 |     11 |  0.0201947511098
           65 |           15796 |       1 |      3 |  0.0440260243272
```

Fig. 3.23 – Similaridade entre documento exemplo e artigo recomendado.

Além do valor da similaridade, também é gravada a posição em que o artigo foi recomendado (1,2 ou 3) e qual o método que foi utilizado.

3.8 Experimentos

O SisRecAC foi avaliado por 35 usuários que realizaram 263 *uploads* e 1614

avaliações. A maioria dos usuários que utilizaram o sistema são alunos da área de computação que foram convidados a utilizá-lo e avaliar os artigos recuperados, segundo a relevância em relação ao documento exemplo. Destaca-se que, quando a recuperação for do próprio documento exemplo, este deveria ser avaliado como “Totalmente Relevante”.

Cada *upload* gera três recomendações que devem ser avaliadas, mas o usuário tem a opção de gerar novas listas com artigos recuperados, podendo cada documento exemplo(*upload)* gerar até vinte novas listas ou sessenta artigos.

3.9 Resultados

Nesta seção serão apresentados os resultados das avaliações subjetiva e objetiva (matemática).

3.9.1 Avaliação Subjetiva

A tabela 3.3 apresenta os resultados obtidos pelos vinte métodos, onde a coluna “Total de Recomendações Avaliadas @1” indica quantas avaliações foram realizadas do primeiro item recomendado e para o método indicado a sua esquerda. Na primeira linha se observa o número 28, ou seja, 28 avaliações do primeiro item recomendado para o método Expr1. A precisão (relevantes + parcialmente relevantes) ficou em 32,14%. Quando somados os três primeiros itens, para o método Expr1, foram avaliados 72 documentos e a precisão alcançou 20,83%.

Com relação ao desempenho, o método PC9, que é a extração das nove palavras mais freqüentes, possui o melhor desempenho, tanto para @1 quanto para @3. Dos métodos que não necessitam de informações extraídas do documento, ou seja, que utilizam informações cadastradas pelo usuário, chamados TagExp1, TagExp2, TagExp3 (tags do usuário + expansão) obtiveram bons resultados, bem superiores ao resultado obtido pelo método Tag (apenas tag), o que demonstra que o uso de expansão baseada em Folksonomia pode ser um recurso promissor em termos de descoberta de conhecimento. Outro método com bom desempenho foi o método Tit, que utiliza as palavras do título (sem *stopwords*). Por outro lado, os métodos que utilizam expressões (Exp1,Exp2,Exp3) obtiveram os piores resultados entre todos os métodos.

Tabela 3.3 – Apresentação dos resultados da avaliação subjetiva.

MÉTODO	Total de Recomendações Avaliadas @1	Precisão para @1	Total de Recomendações Avaliadas @3	Precisão para @3
Expr1	28	32.14 %	72	20.83 %
Expr2	28	25.00 %	75	36.00 %
Expr3	28	57.14 %	69	52.17 %
PC4	32	84.37 %	72	73.61 %
PC5	28	78.57 %	69	72.46 %
PC6	26	88.46 %	72	84.72 %
PC7	30	90.00 %	69	85.50 %
PC8	30	90.00 %	63	92.06 %
PC9	28	96.42 %	75	96.00 %
Tit	29	82.75 %	69	76.81 %
Tag	29	55.17 %	72	65.27 %
TagExp1	28	82.14 %	75	82.66 %
TagExp2	29	82.75 %	78	73.07 %
TagExp3	30	76.66 %	75	80.00 %
PC2Exp2	28	53.57 %	72	54.16 %
PC3Exp3	31	67.74 %	66	75.75 %
PC4Exp4	30	63.33 %	72	70.83 %
PC2ORExp2	33	39.39 %	51	49.01 %
PC3ORExp3	33	78.78 %	60	63.33 %
PC4ORExp4	32	90.62 %	66	80.30 %

Tabela 3.4 – Apresentação dos resultados da avaliação subjetiva @1.

MÉTODO	TOTAL DE AVALIAÇÕES	PRECISÃO MÉDIA (%) @1	
		PARCIALMENTE + RELEVANTE	IRRELEVANTE
Expr1	28	32,15	67,86
Expr2	28	25,00	75,00
Expr3	28	57,14	42,86
PC4	32	84,38	15,63
PC5	28	78,57	21,43
PC6	26	88,46	11,54
PC7	30	90,00	10,00
PC8	30	90,00	10,00
PC9	28	96,43	3,58
Tit	29	82,76	17,25
Tag	29	55,17	44,83
TagExp1	28	82,15	17,86
TagExp2	29	82,76	17,25
TagExp3	30	76,66	23,34
PC2Exp2	28	53,58	46,43
PC3Exp3	31	67,74	32,26
PC4Exp4	30	63,33	36,67
PC2ORExp2	33	39,39	60,61
PC3ORExp3	33	78,79	21,22
PC4ORExp4	32	90,63	9,38

Tabela 3.5 – Apresentação dos resultados da avaliação subjetiva @3.

MÉTODO	TOTAL DE AVALIAÇÕES	PRECISÃO MÉDIA (%) @3	
		PARCIALMENTE + RELEVANTE	IRRELEVANTE
Expr1	72	20,83	79,17
Expr2	75	36,00	64,00
Expr3	69	52,17	47,83
PC4	72	73,61	26,39
PC5	69	72,46	27,54
PC6	72	84,72	15,28
PC7	69	85,50	14,50
PC8	63	92,06	7,94
PC9	75	96,00	4,00
Tit	69	76,81	23,19
Tag	72	65,27	34,73
TagExp1	75	82,66	17,34
TagExp2	78	73,07	26,93
TagExp3	75	80,00	20,00
PC2Exp2	72	54,16	45,84
PC3Exp3	66	75,75	24,25
PC4Exp4	72	70,83	29,17
PC2ORExp2	51	49,01	50,99
PC3ORExp3	60	63,33	36,67
PC4ORExp4	66	80,30	19,70

A fig. 3.24 apresenta um gráfico comparativo da precisão dos vinte métodos e permite visualizar melhor o desempenho destes.

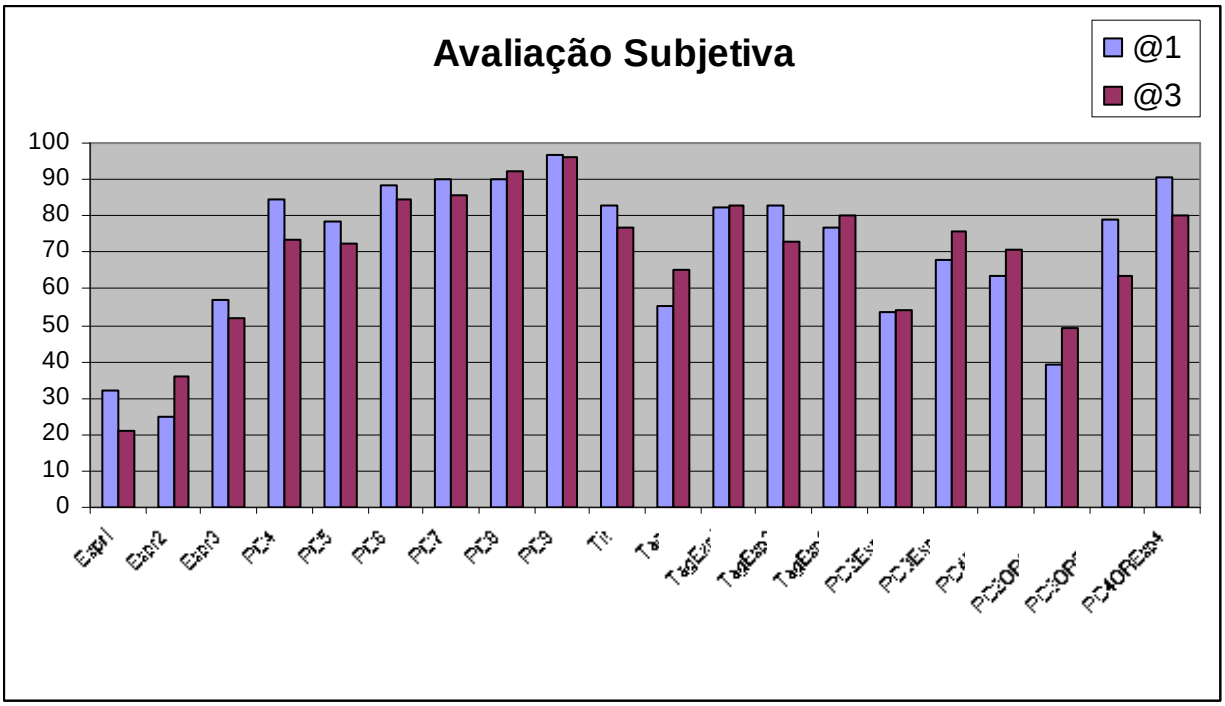

Fig. 3.24 – Gráfico comparativo dos métodos.

O gráfico da figura 3.24 permite visualizar melhor os métodos com desempenho mais eficaz. Por esse gráfico, se observa que o melhor desempenho foi alcançado pelo método que utilizou nove palavras-chave (PC9). Os métodos PC4, PC5, PC6, PC7 e PC8 obtiveram desempenho semelhantes entre si, contudo, abaixo do método PC9. Com relação aos métodos que utilizam as tags dos usuários, o método que utiliza apenas as tags (sem expansão) não obteve uma boa avaliação subjetiva quando comparado aos demais métodos. Por outro lado, quando o método utiliza as tags mais a expansão baseada em Folksonomias, o desempenho melhora sensivelmente, passando de 62,27% sem expansão(Tag em @3) para 82,66% com expansão (TagExp1 em @3), demonstrando que a técnica de expansão tende a melhorar o desempenho dos métodos. Outro método que utiliza informações cadastradas pelos usuários, no caso o título do documento (Tit), obteve desempenho médio (76,81%) e semelhante ao desempenho do método que utiliza tags com expansão.

Os métodos que utilizam expressões alcançaram os piores desempenhos dentre os métodos avaliados, alcançando 20,83%, 36% e 52,17% para os métodos Expr1, Expr2 e Expr3 respectivamente. O algoritmo foi desenvolvido segundo o que está descrito na seção 3.5.1. No escopo deste trabalho não foi realizada uma avaliação em relação à efetividade do algoritmo que gera as expressões. Assim, é possível que a substituição pelo algoritmo EPC-R

ou EPC-P, descritos em Pereira (2001), venha a melhorar o desempenho da avaliação subjetiva em relação aos métodos que utilizam expressões.

Por fim, é importante destacar novamente os métodos PC9, TagExp1 e Tit, que representam os melhores métodos em duas categorias distintas, onde o primeiro representa os métodos que utilizam informações extraídas automaticamente do documento exemplo e o segundo e terceiro utilizam informações cadastradas pelo usuário, no caso as tags mais a expansão baseada em folksonomia (TagExp1) e o título do documento (Tit). A importância desse destaque está no fato de que apesar do método PC9 alcançar um excelente desempenho (96%), em alguns casos tal método talvez não possa ser utilizado pelo SisRecAC para recomendar artigos científicos. Isso ocorre devido à possibilidade de abertura do documento exemplo (formato pdf), o que nem sempre é possível, já que o documento pode estar bloqueado. Nestes casos (inviabilidade de conversão do formato pdf para texto), o melhor método fica entre TagExp1 (82,66%) e Tit (76,81%). A opção por destacar o método Tit como uma segunda opção, dentre aqueles que utilizam informações cadastradas pelo usuário, e não o método TagExp3 (80%) se dá pelo fato de que alguns sistemas optam por deixar o cadastramento das tags de um recurso como opcional. Já em relação ao título, tal falta dificilmente ocorre e a maioria dos sistemas exige o preenchimento desta característica para identificação do recurso.

3.9.2 Avaliação Objetiva (Similaridade)

Para apresentar os resultados utilizou-se a tabela 3.4, onde a primeira coluna da relaciona os métodos e a segunda coluna o total de resultados que foi obtido em cada um dos métodos. A coluna @1, apresenta o valor da similaridade média para o primeiro artigo recuperado e a coluna @3 apresenta a similaridade média para os três primeiros.

A similaridade é calculada entre o artigo recuperado e o documento exemplo (*upload*), ou seja, o documento que gerou a recuperação.

Tabela 3.6 – Resultados da avaliação de similaridade.

Método	Número de Avaliações	SIMILARIDADE MÉDIA	
		@1	@3
Expr1	52	0,1169266939	0,0886325444
Expr2	45	0,1845363267	0,1310112212
Expr3	91	0,2160707058	0,1241280928
PC4	74	0,2742827231	0,1685426444
PC5	84	0,2552811481	0,1715982002
PC6	88	0,2788548730	0,1988855736
PC7	91	0,3001339517	0,1998724846
PC8	97	0,2213364533	0,1738647532
PC9	85	0,2806952282	0,1941831950
Tit	72	0,3410787051	0,2033507406
Tag	74	0,0665717192	0,0738300863
TagExp1	44	0,1204276900	0,0953103231
TagExp2	50	0,1210793233	0,0879749654
TagExp3	48	0,1103482132	0,0909316033
PC2Exp2	35	0,0987418879	0,0751705030
PC3Exp3	47	0,1015108515	0,0810550168
PC4Exp4	49	0,0823812070	0,0721821910
PC2ORExp2	43	0,3165311750	0,1601170551
PC3ORExp3	44	0,2062686022	0,1465791008
PC4ORExp4	40	0,2302052429	0,1243682909

O gráfico 3.24 apresenta os resultados da avaliação de similaridade com duas linhas. Uma representando @1, isto é, a similaridade média (por método) do artigo que aparece em primeiro lugar na lista dos artigos recuperados. Observa-se que em quase todos os métodos a similaridade média de @1 é superior a similaridade média de @3 (conjunto das três recomendações). Somente para o método Tag a similaridade média de @3 superou @1.

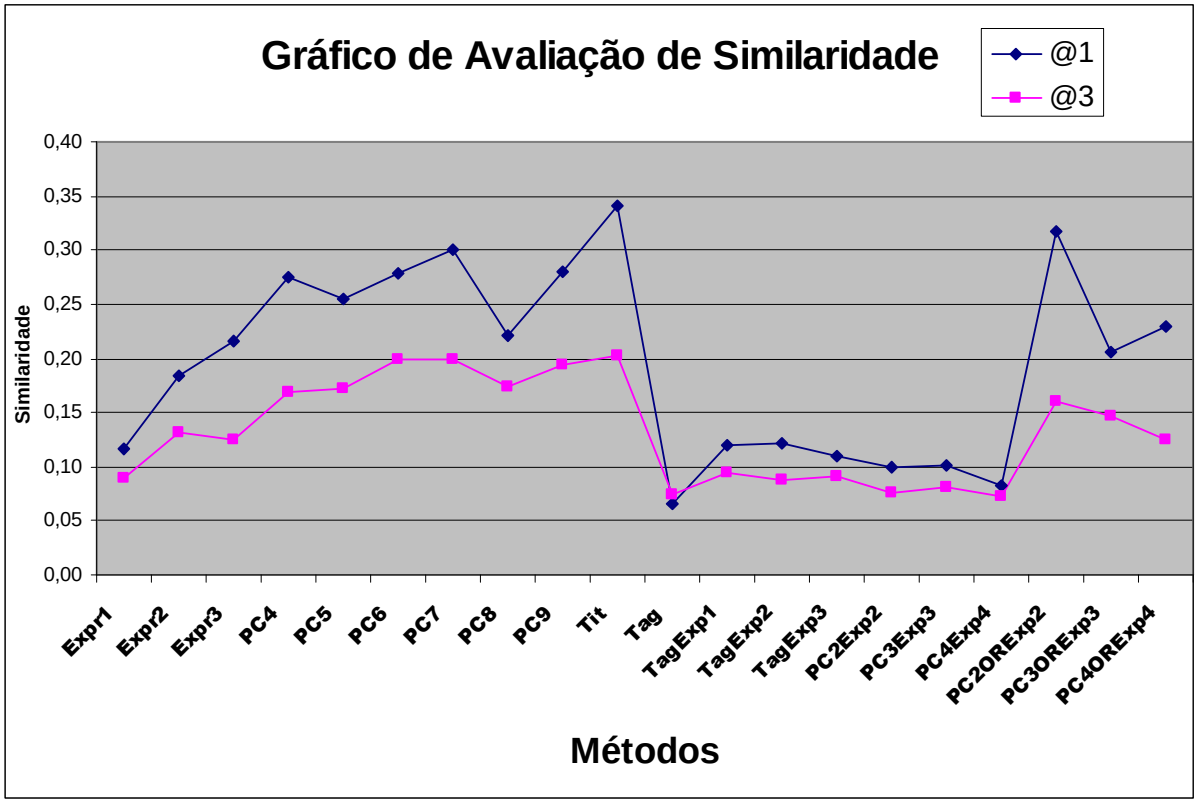

Fig. 3.25 – Gráfico de Avaliação de similaridade (linhas).

No gráfico de colunas (Fig. 3.26) pode ser observado que os métodos que extraem palavras-chave (PC4,PC5,PC6,PC7,PC8 e PC9) e também os métodos cujas palavras-chave são expandidas, utilizando o operador OR (PC2ORExp2, PC3ORExp3 e PC4ORExp4) recuperam artigos mais similares. Quando considerado apenas @1, o método que recuperou artigos mais similares foi o Tit, o qual é baseado nas palavras do título do documento.

Portanto, em termos de similaridade, o melhor avaliado é o método Tit, principalmente considerando a similaridade do documento exemplo com o primeiro artigo recuperado (@1), que alcançou o índice 0,3410787051. Contudo, se for considerada a similaridade média para os três recuperados (@3), o método Tit apresenta desempenho similar aos melhores (PC6, PC7 e PC9).

Dentre os piores métodos em termos de similaridade se destacam os métodos que utilizam as tags cadastradas pelo usuário (Tag), tags mais expansão baseada em Folksonomia e palavras-chave mais expansão baseada em Folksonomia(Tag, TagExp1, TagExp2, TagExp3, PC2Exp2, PC3Exp3, PC4Exp4).

Ainda em relação à similaridade como o documento exemplo, apresentam desempenho intermediário os métodos que utilizam as palavras-chave com expansão baseada em Folksonomia com operador OR (PC2ORExp2,PC3ORExp3,PC4ORExp4) e também os métodos que utilizam expressões (Expr1, Expr2 e Expr3).

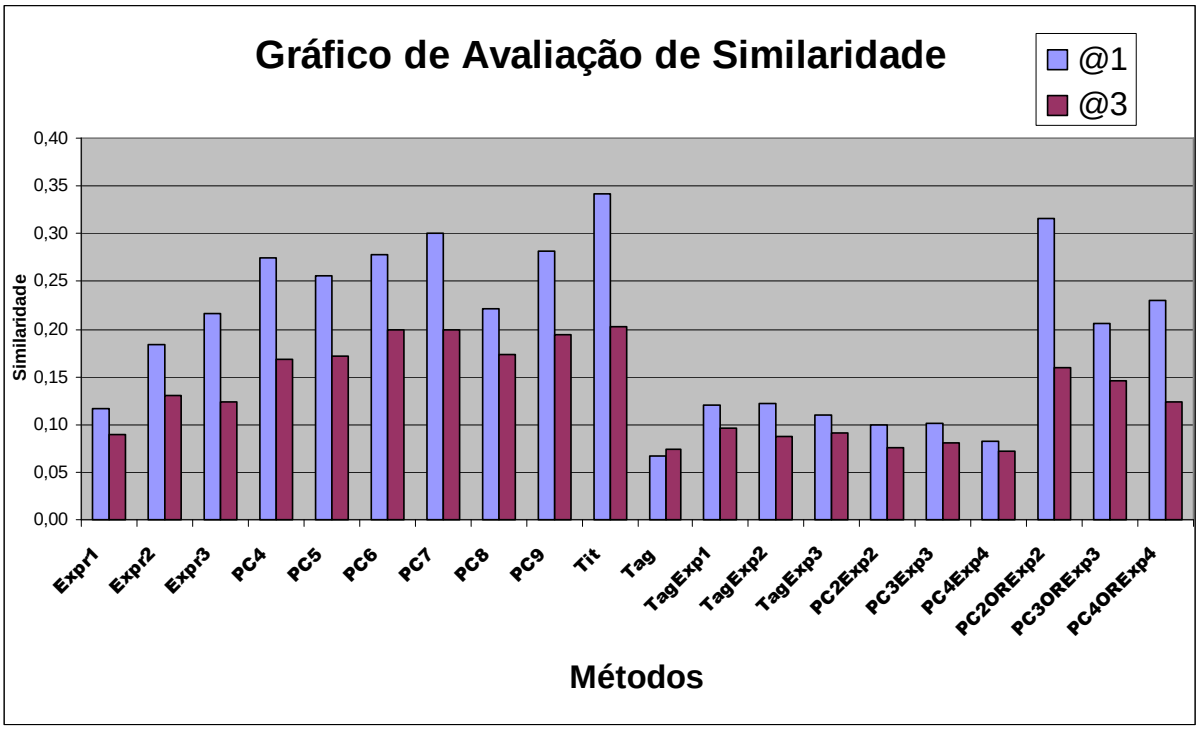

Fig 3.26 - – Gráfico de Avaliação de similaridade (colunas).

No gráfico comparativo (fig. 3.27) entre a avaliação subjetiva (@1) e a avaliação objetiva (@1) (valores foram normalizados – multiplicados por 100) observa-se que ocorreu uma certa correspondência entre relevância e similaridade pois os métodos PC, que são, em geral, os mais similares, obtiveram boa avaliação de relevância por parte dos usuários, assim como o método Tag, que obteve avaliação relativamente baixa de relevância e também baixa similaridade. Um método que apesar da boa similaridade, foi avaliado com baixa relevância foi o PC2ORExp2. Uma comparação entre os métodos que combinam palavras-chave extraídas automaticamente e expandidas com AND (PC2Exp2, PC3Exp3, PC4Exp4) com os métodos que combinam as palavras-chave com expansão ao operador OR (PC2ORExp2, PC3ORExp3, PC4ORExp4) é possível observar que a utilização do operador OR obteve um desempenho bem superior em termos de similaridade.

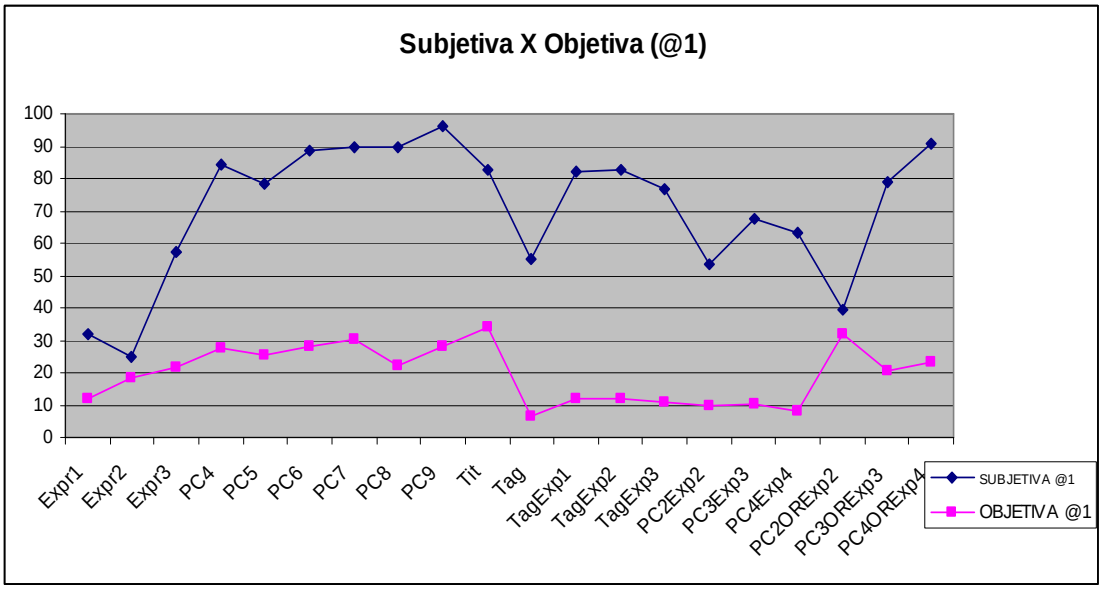

Fig 3.27 – Gráfico comparativo (Avaliação Subjetiva X Avaliação Objetiva @1).

Com relação à comparação para @3 (Fig 3.28), entre a avaliação subjetiva e a objetiva, valem as mesmas considerações de @1, ou seja, a similaridade normalmente acompanha a avaliação subjetiva, também com algumas exceções, como o método PC2ORExp2, que também tem boa similaridade em @3, porém baixa avaliação de relevância.

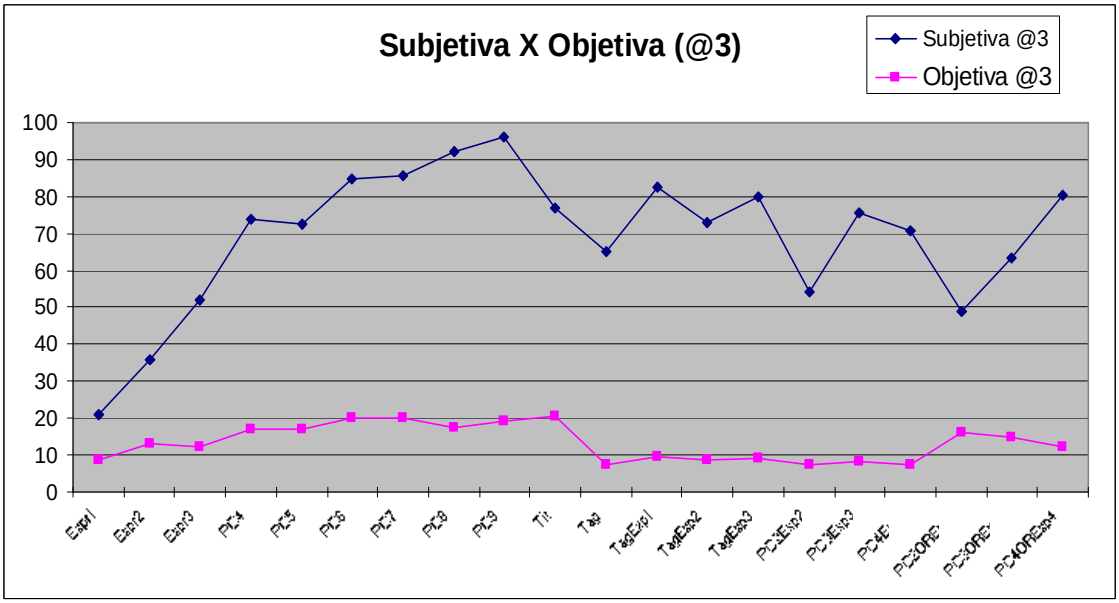

Fig 3.28 – Gráfico comparativo (Avaliação Subjetiva X Avaliação Objetiva @3).

4. Conclusão

O desenvolvimento do SisRecAc, sua utilização por parte de alguns membros da comunidade universitária, as avaliações publicadas em evento de nível nacional e neste documento comprovam que o objetivo do trabalho foi alcançado. Assim, torna-se possível testar diferentes métodos de extração de informações de um documento exemplo para prover aos usuários recomendações de artigos científicos.

Os altos percentuais de relevância alcançados por alguns dos métodos testados comprovam que a utilização de um documento exemplo (QBE), em oposição à digitação de termos para a busca é uma abordagem interessante e que deve ser ainda mais investigada como alternativa para recuperação de informações.

Os resultados da avaliação de similaridade, utilizando a fórmula matemática apresentada em (Loh 2001), com documentos recuperados do Google Acadêmico, a partir de palavras-chave extraídas de documentos mostram que, em geral, a relevância da recuperação acompanha a similaridade do documento. Existem exceções onde, apesar da boa similaridade, a avaliação subjetiva aponta para baixa relevância, como o método Tit, por exemplo, que foi o que recuperou artigos mais similares, porém obteve um percentual de relevância apenas médio quando comparado aos demais. Outra exceção foi o método PC2ORExt2, que também obteve um alto valor de similaridade, porém, baixa avaliação de relevância.

Os resultados demonstram que a busca de informação a partir de termos extraídos automaticamente de um texto exemplo é uma abordagem efetiva para recuperação de informações relevantes. As avaliações subjetivas para o método de nove palavras-chave (método PC9) alcançaram um índice de 96% de relevância ou satisfação do usuário em relação às recomendações. O índice é bem superior ao que foi alcançado com a busca a partir das tags que foram cadastradas pelo usuário para identificar o documento (método tag), que obteve um índice de aproximadamente 65%. Também na avaliação objetiva, os métodos de extração automática de palavras-chave recuperaram, em geral, documentos mais similares do que aqueles recuperados a partir das tags cadastradas pelos usuários. Em uma comparação direta, o índice médio de similaridade (que vai de 0 a 1) para @3 do método PC9 foi de aproximadamente 0,1941, bem superior ao índice de 0,0738 alcançado pelo método tag.

Os resultados obtidos confirmam as conclusões do trabalho de Brooks & Montanez (2006), no qual os autores concluem que quando o processo de identificação do texto é manual, as pessoas, muitas vezes, tendem a identificar o texto com termos que possuem um significado pessoal e não reconhecem o assunto do texto. Da mesma forma, os mesmos autores, concluem que a geração automática de palavras-chave (mais freqüentes) e o conseqüente agrupamento, produzem grupos mais similares (mais focados) do que os grupos gerados a partir da identificação manual (*tagging*).

Outra conclusão importante, baseada nos resultados alcançados, é que o uso de algum tipo de expansão semântica pode melhorar significativamente os resultados. O dado que reforça essa afirmação é a melhoria do percentual de relevância do método tags, que variou de 65%, sem expansão, para 80% com expansão de três termos para cada tag cadastrada pelo usuário. Apesar de o resultado ser bem inferior ao que foi alcançado pelo método PC9 (extração automática), esse método é importante, pois no contexto do SisRecAC, em alguns casos, não será possível extrair as palavras-chave do documento exemplo.

Uma das dificuldades do trabalho foi obter um número considerável de avaliações subjetivas e até mesmo de *uploads*, pois depende dos usuários do sistema. Durante o projeto, principalmente avaliando artigos da área, surgiram idéias para implementação de novos métodos, porém a dificuldade mencionadas fez com que não fosse ampliado ainda mais o número de métodos a serem testados. Uma solução para tal problemática seria a utilização de uma base de referência e apenas a avaliação objetiva (matemática) através da fórmula de similaridade (Loh,2001). Todavia, porém essa abordagem não foi utilizada, já que estudos preliminares detectaram que a avaliação de similaridade nem sempre correspondia à avaliação de relevância.

Todos os resultados apresentados mostram uma tendência em termos de avaliação subjetiva e objetiva, não podendo ser considerada uma verdade absoluta. Tal fato ocorre devido à amostra de usuários, composta, em sua maioria, por alunos de graduação dos cursos de computação da Universidade Católica de Pelotas, sendo possível que no caso de uma amostra com um número maior de usuários, ou ainda de outras áreas do conhecimento, ocorra uma alteração nos resultados.

4.1 Trabalhos Futuros

Um dos métodos ou abordagem que poderia ser testada é aquela baseada em alguma forma de normalização das palavras-chave, como a técnica de redução ao radical, conhecida por *stemming* ou pelo menos a redução do termo ao equivalente no singular e ainda a sua forma no masculino. Métodos baseados em *stemming* poderiam ser testados e comparados com *stemming* mais expansão baseada em Folksonomia

Outro aspecto a destacar é que a ferramenta desenvolvida (SisRecAC) poderá ser acoplada a sistemas de EAD e na medida em que os professores armazenam documentos como conteúdos programáticos, apostilas sobre os assuntos tratados no curso/disciplina, textos nos fóruns de discussão, sessões de chat, entre outros. A ferramenta poderia extrair as palavras-chave e recomendar artigos e bibliografia relevante. Na medida em que o sistema de EAD é utilizado (documentos, fóruns, chat,..) as recomendações vão sendo realizadas e avaliadas pelos professores que poderão recomendar a alunos. Quanto à bibliografia relevante, a idéia é pesquisar e apresentar bibliografia atualizada, foi descoberta nos sites de livrarias digitais e editoras. No que se refere a artigos, é interessante salientar, que não só o Google-Acadêmico seria pesquisado e utilizado como fonte de pesquisa, mas também bibliotecas digitais previamente selecionadas.

Além de sistemas de EAD, o SisRecAC poderia ser integrado a sistemas administrativos, como os de administração hospitalar, por exemplo. Assim, recomendaria artigos científicos baseados em diagnósticos do médico em relação ao paciente. O médico poderia visualizar os artigos no próprio sistema ou através de avisos que poderiam chegar por e-mail.

Além de integrado a outros sistemas, o SisRecAC poderia funcionar também como site/serviço onde o usuário teria a possibilidade de realizar *upload* de arquivos e receber as recomendações baseadas nesses uploads.

Para a implantação do SisRecAC, como ferramenta e não como um instrumento de teste de métodos, será importante definir e utilizar o melhor ou os melhores métodos, de acordo com o caso, pois, ao partir dos resultados desse trabalho, os melhores métodos são baseados na extração de palavras-chave. Destaca-se que essa extração nem sempre é possível, portanto, quando for inviável a extração, poderia ser utilizado o método de utilização das tags

cadastradas pelo usuário mais expansão baseada em Folksonomia.

Durante a realização do trabalho, diversas idéias de melhorias ou de novas funcionalidades foram sugeridas pelos usuários, como por exemplo:

a) Apresentação de mais resultados (definido pelo usuário?) e não apenas três como foi fixado para os experimentos (baseado em Kraft et al, 2006)

b) Recomendação, pela ferramenta, de notícias baseadas nas palavras e frases chaves extraídas do repositório de documentos. Poderia ser realizada uma meta-busca em sites como o http://news.google.com.br, http://br.search.yahoo.com/news.

c) Existência de um módulo para recuperação bibliográfica (livros), também baseado na extração de palavras-chave dos documentos base, retornando os dados do livro e também os locais de venda com preços.

d) Utilização do formato RSS para divulgar as recomendações, desde que autorizado pelo usuário.

e) Integração do SisRecAC com ambientes de EAD livres (Moodle, Claroline, Atutor, Dokeos,...) através da construção de módulos integradores, provavelmente com a utilização de XML.

f) Integração com bibliotecas digitais, onde os usuários poderiam escolher os documentos da biblioteca a serem utilizados como base para a recuperação de artigos científicos.

g) Recomendação para vários amigos, pois atualmente é possível indicar um documento e a lista de documentos localizados para apenas um amigo de cada vez, isto é, apenas um endereço de e-mail de cada vez. A sugestão, neste caso, seria de construir algo semelhante ao oferecido no Google Docs, no qual se pode indicar vários endereços de cada vez. A Figura 4.1 foi extraída da tela de cadastro de colaboração do Google Docs e ilustra a sugestão de melhoria das recomendações.

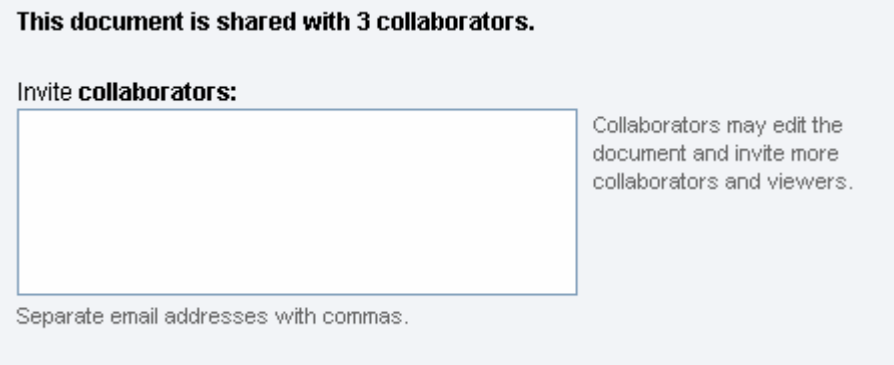

Fig. 4.1 - Google Docs - possibilidade de recomendar para vários e-mails.

h) Validação dos links, tendo em vista que a recuperação do SisRecAc, após a análise do documento exemplo, é apresentada em forma de uma lista de links que apontam para documentos no formato pdf. Verificou-se que diversos links estão quebrados, ou seja, quando o usuário clica é apresentada uma mensagem de erro informando que o documento não foi localizado. Seria interessante a implementação de um mecanismo que verificasse a integridade dos links excluindo da lista aqueles que não estivessem disponíveis.

i) Conversão de um documento localizado em documento exemplo para novas recomendações, ou seja, disponibilizar a opção de transformar um documento que foi recomendado pelo sistema em documento exemplo para gerar novas recomendações. Atualmente, essa operação requer que o usuário faça o *download* do documento e posteriormente faça o *upload* para o servidor.

j) Implementação de outros tipos de documentos e digitação do documento exemplo, pois atualmente é possível submeter ao SisRecAC apenas documentos do tipo texto ou pdf. Seria interessante a possibilidade de envio de outros tipos, como.doc (padrão Microsoft Word), ppt (Microsoft Powerpoint) ou ainda as extensões utilizadas pelo OpenOffice. Outra possibilidade é permitir que o usuário digite o texto (documento exemplo) diretamente no sistema. Poderia ser disponibilizado no

SisRecAC um editor como o FCKeditor[26] por exemplo. Dessa forma, um autor poderia estar redigindo um artigo científico e de forma dinâmica e periódica, artigos científicos estariam sendo recomendados.

k) Destaque de documentos que são localizados a partir de mais de um documento, pois verificou-se que durante a utilização do SisRecAC alguns documentos eram recomendados para mais de um documento exemplo. Seria interessante destacar esse fato, identificando o documento e a relação de documentos base que o localizou.

l) Desenvolvimento de um sistema de Agentes para monitorar novos documentos

Uma funcionalidade extra, que não constava no projeto inicial, foi a implementação no SisRecAC do conceito de Agentes de Software. Nesse caso, os agentes desenvolvidos são responsáveis pela descoberta, armazenamento e envio de avisos sobre novos artigos que não constavam na lista inicial, recomendada pelo sistema quando do *upload* do documento exemplo. Os usuários que desejam, podem receber periodicamente, por e-mail, uma listagem das novas recomendações (novos artigos descobertos, tendo como base um documento que foi enviado).

[26] http://www.fckeditor.net/

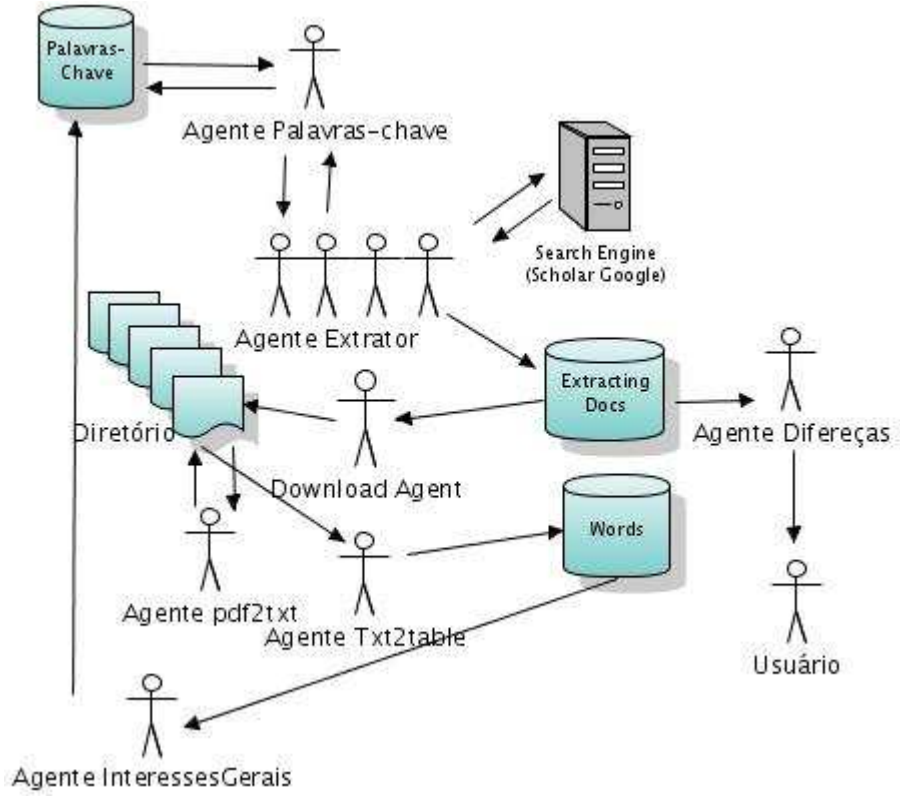

Fig. 4.2 – Relação entre os agentes.

A figura 4.2 mostra a relação entre os diversos agentes que foram desenvolvidos para gerar as recomendações periódicas.

A figura 4.3 apresenta o e-mail recebido pelo usuário, enviado pelo agente Avisos, com a listagem das diferenças localizadas pelo agente Diferenças, com a colaboração dos demais agentes.

Fig. 4.3 - E-mail enviado pelo agente Avisos

Trabalhos Publicados

ÁVILA, Christiano Otero ; LOH, S. . SisRecAC - Sistema de Recomendação de Artigos Científicos. In: XII Simpósio Brasileiro de Sistemas Multimídia e Web - WEBMEDIA (ferramentas), 2006, Natal. Anais XII Simpósio Brasileiro de Sistemas Multimídia e Web -

WEBMEDIA, 2006.

ÁVILA, Christiano Otero ; Silva, Rosaura ; PALAZZO, Luiz Antônio Moro ; LOH, S. . Utilização de Sistemas Multiagentes na Construção de Sistemas de Recomendação. In: Workshop Escola de Sistemas de Agentes para Ambientes Colaborativos, 2007, Pelotas. Anais do Workshop Escola de Sistemas de Agentes para Ambientes Colaborativos, 2007.

ÁVILA, Christiano Otero ; LOH, S. ; FONSECA, Frederico da Rocha . Sistema de Recomendação de Artigos Científicos a Partir de um Texto Exemplo. In: XIII Webmedia - Brazilian Symposium on Multimedia and the Web, 2007, Gramado, RS. Anais do XIII Simpósio Brasileiro de Multimídia e Web, 2007. p. 214-221

Bibliografia

ADOMAVICIUS, G.; TUZHILIN, A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering, v.17 n.6, p.734-749, Junho 2005

ADRIAN, B.; SAUERMANN, L.; ROTH-BERGHOFER, T. Contag: A semantic tag recommendation system. Proceedings of I-Semantics' 2007. JUCS, p. 297-304. URL http://www.dfki.uni-kl.de/~sauermann/papers/adrian+2007a.pdf

AIRES, Rachel ; ALUÍSIO, S. M.; KUHN, D. C. S.; ANDREETA, M. L. B.; Oliveira Jr., O. N. Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. (SBIA'2000) Atibaia, SP, November, 20−22, 2000.

BEGELMAN, G.; KELLER, P. and SMADJA, F. Automated Tag Clustering: Improving search and exploration in the tag space. WWW2006, Maio 22–26, 2006, Edinburgh, UK.

BELKIN, N. J.; ODDY, R. N.; BROOKS, H. M. ASK for information retrieval: part I. background and theory. In: [SPA97]

BILLERBECK, B.; SCHOLER, F.; WILLIAMS, H. E.; ZOBEL, J. Query expansion using associated queries, Proceedings of the twelfth international conference on Information and knowledge management, Novembro 03-08, 2003, New Orleans, LA, USA

BILLERBECK, B.; ZOBEL, J. Questioning query expansion: an examination of behaviour and parameters, Proceedings of the fifteenth Australasian database conference, p.69-76, Janeiro 01, 2004, Dunedin, New Zealand

BRAUN, S.; SCHMIDT, A.; WALTER, A.; ZACHARIAS, V. The Ontology Maturing Approach for Collaborative and Work Integrated Ontology Development: Evaluation Results and Future Directions FZI Research Center for Information Technologies Information Process Engineering Haid-und-Neu-Straße 10-14, 76131 Karlsruhe, Germany.

BROOKS, C. H.; MONTANEZ, N. Improved annotation of the blogosphere via autotagging

and hierarchical clustering. In: International World Wide Web Conference – WWW, Maio 2006, Edinburgh, Scotland, p.625-631.

CHEKURI, C.; GOLDWASSER, M.; RAGHAVAN, P.; UPFAL, E. Web Search Using Automated Classification. In 6th InternationalWorldWideWeb Conference, (Poster no. POS725), Santa Clara, California, April.

DAHLEN,B.J.; KONSTAN,J.A.; HERLOCKER,J.L.; GOOD,N.; BORCHERS,A.; RIEDL,J. Jump-starting movielens: User benefits of starting a collaborative filtering system with "dead data". University of Minnesota TR 98-017.

DALAL, M. Personalized Social & Real-Time Collaborative Search. Poster on WWW 2007, Maio 8–12, 2007, Banff, Alberta, Canada. URL www2007.org/posters/poster887.pdf .

DAMME, C., HEPP, M.; SIORPAES, K. FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies, Bridging the Gap between Semantic Web and Web 2.0 Workshop at the ESWC 2007. 2007, Innsbruck, Austria.

DIAS, M. A. L. Extração Automática de Palavras-Chave na Língua Portuguesa Aplicada a Dissertações e Teses da Área das Engenharias. Dissertação de Mestrado. Campinas: FEEC-UNICAMP, 2004.

ERRAMI, M; WREN, JD; HICKS, J; GARNER, H. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res. (2007) 35:W12–W15.

FERRAGINA, P. ; GULLI, A. (2005) A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. In the Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, ISBN:1-59593-051-5. 801-810.

FONSECA, Bruno; GOLGHER, Paulo ; PÔSSAS, Bruno , RIBEIRO-NETO, Berthier; ZIVIANI , Nivio. Concept-based interactive query expansion, Proceedings of the 14th ACM international conference on Information and knowledge management, October 31-November 05, 2005, Bremen, Germany [doi>10.1145/1099554.1099726]

GOLDBERG, D.; OKI, B.; NICHOLS, David;TERRY, D. Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 12 (Dec.1992), 61—70.

HERLOCKER, J., KONSTAN, J., TERVEEN, L.; RIEDL, J. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22, 1 (2004), 5–53

IPROSPECT. Search Engine Marketing Firm iProspect Study Reveals Increase in the Importance of Attaining Top Natural Search Results. Disponível em:

< http://www.iprospect.com/media/press2006_04_11.htm > . Acesso em: Dezembro 2006.

KRAFT, R.; CHANG, C.; MAGHOUL, F. ; KUMAR, R. 2006. Searching with context. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 477-486.

LAU, Tessa ; HORVITZ, Eric (1999) Patterns of search: analyzing and modeling web query refinement. In: 7th International Conference on User Modeling, June 1999, Banff, Canada, p.119-128

LEE, L.. Measures of distributional similarity. In Proceedings of the 37th ACL, 1999

LEWIS,J.; OSSOWSKI,S.; HICKS,J.; ERRAMI,M.; GARNER,H. (2006) Text similarity: an alternative way to search medline.Bioinformatics, 22, 2298–2304.

LOH, S. Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos. Porto Alegre: UFRGS. Requisito Parcial ao Grau de Doutor em Ciência da Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, 2001.

MATHES, A. (2004). Folksonomies – Cooperative Classification and Communication Through Shared Metadata.
http://www.adammathes.com/academic/computer-ediatedcommunication/folksonomies.html

MCNEE, S. ; RIEDL, J. ; KONSTAN, J. . Accurate is not always good: How accuracy metrics have hurt recommender systems. ACM CHI 2006.

ORENGO, V. ; HUYCK, C. Stemming Algorithim for The Portuguese Language. In: Proceedings of the SPIRE Conference. Laguna de San Raphael: [s.n.], 2001, p. 13-15.

OSINSKI S.; STEFANOWSKI J.; WEISS, D. Lingo: Search results clustering algorithm based on Singular Value Decomposition. Submitted to Intelligent Information Systems Conference 2004, Zakopane, Poland, 2003.

PARALIC, J.; KOSTIAL, I. Ontology-based Information Retrieval, In Proceedings of the 14th International Conference on Information and Intelligent Systems, ISBN 953-6071-22-3, pp 23-28, 2003.

PEREIRA, M, SOUZA, C; NUNES, M. Implementação, Avaliação e Validação de Algoritmos de Extração de Palavras-Chave de Textos Científicos em Português. SBC - Revista Eletrônica de Iniciação Científica, 2002

PORTER, M.F., 1980. “An Algorithm for Suffix Stripping. Program”, vol.14, n. 3, pp. 130-137.

POTHEN, A.;SIMON,H; LIOU, Kan-Pu. Partitioning sparse matrices with eigenvectors of graphs. SIAM J.Matrix Anal. Appl., 11(3):430{452, 1990.

PRUD’HOMMEAUX, E. , SEABORNE, A., “SPARQL Query Language for RDF, W3C Candidate Recommendation 6 April 2006.” http://www.w3.org/TR/2006/CR-rdf-sparql-query-20060406/ (viewed 25/07/06), 2006.

RESNICK, P. ; VARIAN, H. R. 1997. Recommender systems. Commun. ACM 40, 56–58.

RIBEIRO-NETO, B; Cristo, M.; MOURA, E. S. de; GOLGHER, P. B. Impedance coupling in content-target advertising. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 496--500, Salvador, Bahia, Brasil, Julho 2005.

SALTON, G.; MCGILL, M. J. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

SANDERSON, M. ;CROFT, B. (1999) “Deriving concept hierarchies from text” In: Proceedings of the 22nd ACM Conference of the Special Interest Group in Information

Retrieval, pp. 206-213.

SCHAFER, J. ; KONSTAN, J; RIEDL, J. (2001) E-commerce recommendation applications. Journal of Data Mining and Knowledge Discovery, v.5, n.1/2, Janeiro, p.115-153.

SCHMITZ, C; HOTHO, A ; ASCHKE, R J; STUMME,G. Mining Association Rules in Folksonomies. In Proceedings of the 10th IFCS Conference, 2006.

SILVERSTEIN, C.; HENZINGER, M.; MARAIS, H.; MORICZ, M. (1999) Analysis of a very large web search engine query log. ACM SIGIR Forum, 1999, v.33, n.3, p.6-12.

SIORPAES, K., HEPP, M., KLOTZ, A. ; WALTL, M. myOntology: Tapping the Wisdom of Crowds for Building Ontologies, in Poster and Demo Proceedings of the 6th International Semantic Web Conference (ISWC 2007). Busan, Korea, November 11-15, 2007. Poster (pdf)

SIORPAES, K. Towards Using Wikis for Collaborative and Community-Driven Ontology Engineering, accepted for the Doctoral Consortium at ISWC 2007 in Proceedings 6th International Semantic Web Conference (ISWC 2007B) , Springer LNCS, November 11, 2007, Busan, Korea. (forthcoming)

SMITH, G. (2004) “Folksonomy: social classification.” August, 2004. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html

SPINK, Amanda; WOLFRAM, Dietmar; JANSEN, Major B. J.; SARACEVIC, Tefko (2001) Searching the web: the public and their queries. Journal of the American Society for Information Science and Technology, v.52, n.3, p.226 – 234.

SPECIA, L. ; MOTTA, E.: Integrating Folksonomies with the Semantic Web, in: Proceedings of the European Semantic Web Conference (ESWC 2007), Innsbruck, Austria: Springer, 2007.

SOOD , S.; OWSLEY, S.; HAMMOND, K. ; BIRNBAUM, L. Tagassist: Automatic tag suggestion for blog posts. March 2007.

TEEVAN, J.; ADAR, Eytan; JONES, R.; POTTS, M. (2006) History repeats itself: repeat queries in Yahoo’s logs. In: Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR, p.703-704.

TURNEY, P. Learning to Extract Keyphrases from Text, Tech. Report Number NRC−41622, National Research Council Canada, Institute for Information Technology, 1999.

WEISS, Dawid; STEFANOWSKI, J. Web search results clustering in Polish: Experimental evaluation of Carrot. In Proceedings of the New Trends in Intelligent Information Processing and Web Mining Conference, Zakopane, Poland, 2003.

WENGER, E. (1998). Communities of Practice – learning, meaning and identity. Cambridge: Cambridge University Press. WENGER E (1998) Communities of Practice: Learning, Meaning, and Identity. Cambridge University Press, Cambridge.

WIKIPÉDIA. Desenvolvido pela Wikimedia Foundation. Apresenta conteúdo enciclopédico. Disponível em: <http://pt.wikipedia.org/w/index.php?title=Blogosfera&oldid=9407119>. Acesso em: 19 Fev 2008

WITTEN, I. et al. KEA: Practical automatic keyphrase extraction. In:Proceedings of the Fourth ACM Conference on Digital Libraries. [S.l.]: [s.n.], 1999. p. 254-255.

WU, H.; ZUBAIR, M.; MALY, K. 2006. Harvesting social knowledge from folksonomies. In Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark, August 22 - 25, 2006). HYPERTEXT '06. ACM Press, New York, NY, 111-114. DOI= http://doi.acm.org/10.1145/1149941.1149962

YANG, Yiming; PEDERSEN, Jan. A Comparative Study on Feature Selection in Text Categorization, Proceedings of the Fourteenth International Conference on Machine Learning, p.412-420, July 08-12, 1997

XU,Jinxi; Croft, W. Query expansion using local and global document analysis, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.4-11, August 18-22, 1996, Zurich, Switzerland [doi>10.1145/243199.243202]

ZHANG, Y.; CALLAN J.; MINKA, T. “Novelty and Redundancy Detection in Adaptive Filtering,” Proc. 25th Ann. Int’l ACM SIGIR Conf., pp. 81-88, 2002.

ZACHARIAS, V.; BRAUN, S.: Soboleo - Social Bookmarking and Lighweight Engineering of Ontologies. In: Proceedings of theWorkshop on Social and Collaborative Construction of

Structured Knowledge at 16th International World Wide Web Conference (WWW2007). (2007)

ZENG, H.; HE, Q.; CHEN, Z. & Ma, W.(2004). Learning to cluster web search results.

In the Proceedings of the 27th annual international conference on Research and development in information retrieval, Sheffield, UK, ISBN:1-58113-881-4, 210-217.

ZIEGLER, C.N.; MCNEE, S.M.; KONSTAN, J.A.; LAUSEN, G. Improving Recommendation Lists through Topic Diversification. In Proc. of WWW 2005, ACM Press (2005), 22-32.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo