UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FILIPE ROSEIRO CÔGO

Uma abordagem para o uso do conceito de Folkauthority em sistemas de recuperação de informação

Maringá
2012

FILIPE ROSEIRO CÔGO

Uma abordagem para o uso do conceito de Folkauthority em sistemas de recuperação de informação

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Sérgio Roberto Pereira da Silva

Maringá
2012

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

C676a

Côgo, Filipe Roseiro
Uma abordagem para o uso do conceito de Folkauthority em sistemas de recuperação de informação / Filipe Roseiro Côgo. -- Maringá, 2012.
140 f. : il. col., figs., tabs., algs.

Orientador: Prof. Dr. Sérgio Roberto Pereira da Silva.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2012.

1. Recuperação de informação. 2. Folkauthority. 3. Recuperação social de informação. 4. Redes sociais on-line. 5. Folksonomia. 6. Autoridade cognitiva. I. Silva, Sérgio Roberto Pereira da, orient. II. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 21.ed. 025.04
AMMA-00334

FOLHA DE APROVAÇÃO

FILIPE ROSEIRO CÔGO

Uma abordagem para uso do conceito Folkauthority em sistemas de recuperação de informação

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Sérgio Roberto Pereira da Silva
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Itana Maria de Souza Gimenes
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Cesar Augusto Tacla
Universidade Tecnológica Federal do Paraná – CPGEI/UTFPR

Aprovada em: 29 de fevereiro de 2012.
Local da defesa: Sala 007, Bloco C56, *campus* da Universidade Estadual de Maringá

Uma abordagem para o uso do conceito de Folkauthority em sistemas de recuperação de informação

RESUMO

Com a sobrecarga de informação existente no ambiente da *Web* 2.0, a tarefa de se recuperar informação de qualidade e de relevância tornou-se bastante árdua, uma vez que os usuários não são capazes de examinar e interpretar uma parte significativa da informação disponível. Vários autores relacionam a qualidade e a relevância das informações disponíveis na *Web* com o conceito de *autoridade cognitiva*, afirmando que as entidades possuidoras de conhecimento em determinado assunto tendem a publicar informação contendo essas duas características. Assim sendo, este trabalho considera a hipótese de que a utilização de um esquema de *ranking* que leve em conta a autoridade cognitiva das fontes de informação produz um Sistema de Recuperação de Informação que apresenta resultados de busca contendo informação de maior qualidade e relevância aos usuários. Para verificar essa hipótese foi adotada a abordagem denominada de *Folkauthority* – um arcabouço no qual as fontes de informação de um sistema são categorizadas por meio de *tags* de acordo com suas autoridades cognitivas – e proposta uma arquitetura para recuperação de informação cujo esquema de *ranking*, denominado de *AuthorityRank*, baseia-se na categorização das autoridades. As concessões de autoridade cognitivas entre os usuários foram simuladas com base em dados de uma rede social real, sendo implementado um Sistema de Recuperação de Informação cujo esquema de *ranking* considera o arcabouço estabelecido pela abordagem de *Folkauthority*. Esse esquema foi então comparado com outros esquemas utilizando-se a métrica de NDCG, sendo possível verificar a partir de uma análise dos dados resultantes uma melhoria estatisticamente significativa na qualidade e na relevância das informações recuperadas por meio do esquema *AuthorityRank* quando comparado com o esquema *tf-idf*, confirmando, assim, a validade da hipótese com relação à abordagem proposta.

Palavras-chave: Recuperação de Informação. Autoridade Cognitiva. Recuperação Social de Informação. Redes Sociais. Folksonomias. Folkauthority.

An approach for the use of the concept of Folkauthority in information retrieval systems

ABSTRACT

The information overload in the Web 2.0 environment makes the task of retrieving information with quality and relevance quite difficult, since users are not able to examine and interpret a significant part of the information available. Several authors relate the quality and relevance of information available on the Web with the concept of *cognitive authority* stating that entities possessing expertise in particular subject tend to publish information containing these two features. Therefore, this dissertation considers the hypothesis that the use of a ranking scheme that takes into account the cognitive authority of information sources produces an Information Retrieval System that presents search results containing information of higher quality and relevance to users. Was adopted an approach called Folkauthority to verify this hypothesis – a framework in which the sources of information in the system are categorized with tags according to their cognitive authority – and proposed an architecture for information retrieval whose ranking scheme, called AuthorityRank, is based on the categorization of authorities. The concessions of authority among users were simulated based on data from a real social network and an Information Retrieval System whose ranking scheme considers the framework established by the Folkauthority approach was implemented. This scheme was then compared with other schemes using the metric of NDCG, by which was possible to verify a statistically significant improvement in the quality and relevance of information retrieved through the AuthorityRank scheme when compared with the tf-idf scheme, thus, confirming the validity of the hypothesis with respect to the proposed approach.

***Keywords*:** Information retrieval. Cognitive Authority. Social Information Retrieval. Social Network. Folksonomy. Folkauthority.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SBF	Sistema Baseado em Folksonomia
RI	Recuperação de Informação
RSI	Recuperação Social de Informação
SRI	Sistema de Recuperação de Informação
SSRI	Sistema Social de Recuperação de Informação
TF	Term Frequency
IDF	Inverse Document Frequency
DCG	Discounted Cumulated Gain
NDCG	Normalized Discounted Cumulated Gain
AR	AuthorityRank
QI	Qualidade da Informação
URL	Universal Resource Locator

SUMÁRIO

Capítulo

1

Introdução

O ambiente da *Web* e a consolidação do conceito de *Web 2.0* (Murugesan, 2007; O'Reilly, 2007) permitiram a produção e a publicação de informação de uma maneira jamais vista, na qual qualquer pessoa pode produzir e/ou publicar informação sem nenhum tipo de controle ou verificação da qualidade. Como consequência, a atual produção de informação vem ocorrendo em uma escala cuja dimensão gera um efeito chamado de *sobrecarga de informação* (Himma, 2007; Vlahovic, 2011), tornando os mecanismos para a organização de informação bastante necessários. Originalmente, esta questão foi tratada basicamente com duas abordagens: i) a utilização de esquemas de classificação/categorização, os quais muitas vezes necessitam de especialistas para sua elaboração, e ii) a utilização de mecanismos de busca, os quais são baseados na indexação automática dos documentos. Atualmente, os mecanismos de busca representam a maior parte dos sistemas utilizados na *Web*, no entanto eles ainda não são capazes de tratar a questão da filtragem da grande quantidade de informação de forma eficaz.

Nesse contexto, na última década, uma terceira abordagem para o problema da sobrecarga de informação emergiu, a qual combina a abordagem de categorização com a de indexação automática. Esta técnica é denominada de *tagging* e caracterizada pelo fato dos termos (*tags*) que descrevem as informações em um Sistema de Recuperação de Informação (SRI) serem gerados pelos próprios usuários e as buscas serem realizadas sobre esses termos. Aplicações *Web* que utilizam a abordagem de *tagging* para categorizar informações produzidas pelos usuários vêm crescendo em popularidade. Exemplos destas

aplicações incluem os sistemas *Delicious*[1], *Flickr*[2] e *Technorati*[3]. Tais sistemas empregam a abordagem de Folksonomia, a qual se baseia na abordagem de *tagging*, representando uma forma colaborativa de se indexar manualmente as informações contidas em um sistema (Voss, 2007).

Como em um Sistema Baseado em Folksonomia (SBF) o processo de categorização[4] (Rosch, 1978; Trant, 2009) dos documentos disponibilizados é realizado pelos próprios usuários, a qualidade do resultado desse processo muitas vezes depende de *quem* o realizou. Essa afirmação é válida, principalmente, em contextos cujo escopo seja limitado, de forma que se tenha um entendimento compartilhado a respeito do vocabulário e da definição dos itens de informação. Assim, alguns usuários podem ser melhores (ou piores) do que outros para desenvolver esquemas de categorização, bem como para serem utilizados como fonte de informação em determinado assunto (Pereira e da Silva, 2008b). Isso significa que, para obter bons resultados em termos de organização e indexação dos documentos, o processo de categorização nos SBFs depende do conhecimento e das competências de quem o realiza, trazendo a tona a questão da definição da autoridade das fontes de informação em um SBF.

Por outro lado, a área de Recuperação de Informação (RI) passou a interagir com diversas outras áreas, tais como o Design de Interação (Hearst, 2009) e a Análise de Redes Sociais (Goh et al., 2007), na tentativa de prover SRIs que pudessem lidar de forma eficaz e eficiente com a tarefa de encontrar informação na *Web* (Krause et al., 2008; Lachica et al., 2008; Vlahovic, 2011; Zhou et al., 2008). Em especial, a interação com a área de Redes Sociais tem levado a um campo de estudo na área de RI cujas oportunidades e desafios têm sido bastante discutidos nos últimos anos (Chen e Zhang, 2009; Eklund et al., 2010; Hotho et al., 2006; Krause et al., 2008; Voss, 2007; Yong, 2011), gerando uma área denominada de Recuperação Social de Informação (RSI), a qual descreve abordagens e técnicas para a RI em ambientes nos quais é possível considerar a relação social entre os usuários para a busca e o *ranking* de informação (Chevalier et al., 2008; Goh et al., 2007).

Nesse contexto, a abordagem chamada de *Folkauthority*, proposta em Pereira (2008), propõe inovar no tratamento do problema de autoridade pela aplicação da abordagem da Folksonomia na definição das autoridades das fontes de informação. Nessa abordagem, considera-se que cada usuário seja responsável por publicar informações, isto é, cada usuário é considerado uma *fonte de informação* para o sistema. Além disso, permite-se que

[1]http://delicious.com/
[2]http://www.flickr.com
[3]http://technorati.com/
[4]Neste trabalho, o processo de categorização refere-se à atribuição de *tags* aos documentos com o objetivo de descrever o significado do mesmo.

sejam aplicadas *tags* a estes usuários, de forma a categorizar cada fonte de informação de acordo com suas autoridades cognitivas. Assim, o propósito de atribuir *tags* às entidades fontes de informação não será descrever o que essa entidade "é", mas sim identificar o que essa entidade "sabe", ou em que assunto ela é digna de confiança na opinião de quem atribui a *tag* (Pereira e da Silva, 2008b).

A abordagem de *Folkauthority* possui justificativas teóricas embasadas na teoria de Wilson (1983) sobre *autoridade cognitiva*, a qual explica o tipo de autoridade que as pessoas concedem a uma entidade possuidora de conhecimento útil em determinado tópico. A autoridade cognitiva representa a influência que uma entidade pode causar no pensamento de outro indivíduo, de forma a definir "quem sabe o quê sobre o quê". A teoria de Wilson sobre autoridade cognitiva está relacionada ao processo de aquisição de conhecimento de "segunda mão", no qual uma pessoa, a fim de adquirir informação em determinado assunto, realiza uma consulta a uma entidade[5] possuidora de conhecimento, denominada de autoridade cognitiva. Este processo diferencia-se do conhecimento de "primeira mão", no qual o conhecimento é adquirido a partir da experiência ou da verificação própria e não por intermédio de uma entidade.

A discussão sobre autoridade cognitiva também relaciona-se ao controle da qualidade das informações adquiridas/recuperadas. A Qualidade da Informação (QI) é um conceito que refere-se às diversas dimensões pelas quais uma informação pode ser avaliada, as quais incluem o valor, a completude, a validade, a confiabilidade, a credibilidade e a autoridade de uma informação, dentre outras (Rieh, 2002). Existe uma vasta literatura relacionando o tema de autoridade cognitiva com o conceito de QI, principalmente no contexto da *Web* (Fritch e Cromwell, 2001; McKenzie, 2003; Metzger, 2007; Rieh, 2002; Savolainen, 2007; Wathen e Burkell, 2002), sugerindo que o conceito de autoridade cognitiva pode ser utilizado no projeto de um SRI com a finalidade de auxiliar na tarefa de se recuperar informação com o máximo de qualidade e relevância.

Uma vez que no arcabouço estabelecido pela abordagem de *Folkauthority* as fontes de informação são deliberadamente categorizadas pelos usuários de acordo com suas autoridades cognitivas, considera-se que é possível priorizar uma informação em determinado assunto no momento de se calcular o *ranking* dessa informação em uma consulta, utilizando para tal a relação social de concessão de autoridades cognitivas entre os usuários (Pereira e da Silva, 2008b). Com base nessa questão, neste trabalho é levantada a hipótese de que a consideração explícita das autoridades cognitivas por um esquema de

[5]Segundo Wilson (1983) uma *entidade* pode ser uma pessoa, um livro ou uma instituição. É possível citar o exemplo de Wilson (1983) sobre o dicionário, o qual costuma ser considerado uma autoridade cognitiva quando trata-se da definição de termos em uma língua.

ranking pode fazer com que os primeiros resultados obtidos em uma busca possuam maior qualidade e relevância. Dessa forma, esta pesquisa tem o objetivo geral de avaliar como a abordagem de *Folkauthority* pode ser utilizada no contexto da RI em SBFs e quais os efeitos dessa abordagem sobre a qualidade e a relevância das informações recuperadas. Para atingir esses objetivos, é necessário: i) definir uma abordagem para a RI que considere a influência das relações de concessão de autoridades cognitivas entre os usuários do SRI, ii) especificar um esquema de *ranking* para as informações recuperadas em um sistema que utilize a abordagem de *Folkauthority*, e iii) avaliar os resultados gerados a partir do esquema de *ranking* especificado.

Em termos de pesquisas sobre a influência da concessão de autoridades cognitivas para a busca na *Web*, existem alguns poucos trabalhos que estudaram aspectos da utilização do conceito de autoridade cognitiva em SBFs como, por exemplo, o trabalho de Russel (2005), o qual foi seminal em discutir e apresentar a concessão de autoridade cognitiva por meio da técnica de Folksonomia e introduzir um sistema para a identificação de autoridades em determinado assunto utilizando *tags*. No entanto, nesse trabalho não foi discutido como a atividade de concessão de autoridades cognitivas em SBFs pode ser modelada para fins de RI, nem tampouco se tal atividade pode ser benéfica para a tarefa de RI. Em Pereira e da Silva (2008c), a utilização da abordagem de *Folkauthority* foi estudada sob o ponto de vista da redução na sobrecarga de informação em um SRI, tendo sido demonstrados vários benefícios com relação à utilização dessa abordagem. No entanto, é possível indicar algumas restrições quanto a esse trabalho no que diz respeito aos procedimentos utilizados. Os processos de concessão de autoridades cognitivas por meio de *tags* e de categorização das informações foram definidos de forma aleatória, divergindo daquilo que é apontado na literatura como um modelo para descrever esses processos (Golder e Huberman, 2006; Kleinberg, 1999; Mika, 2007). Além disso, em Pereira e da Silva (2008c) os resultados da utilização da abordagem de *Folkauthority* para RI foram comparados com uma abordagem para RI na qual os resultados eram apresentados em ordem cronológica de publicação e não de acordo com um esquema de *ranking* tradicional para a RI na *Web*.

Neste trabalho, é proposto o uso de informações extraídas da rede social denominada de *cadeia de autoridades* (Pereira e da Silva, 2008c), a qual é gerada a partir da atividade de concessão de autoridades cognitivas dentro do arcabouço estabelecido pela abordagem de *Folkauthority*. A cadeia de autoridades revela com quais *tags* e com quais níveis as autoridades foram categorizadas e pode ser modelada a partir de um grafo rotulado orientado ponderado, no qual uma aresta orientada do nó A ao nó B indica uma concessão de autoridade *do usuário* representado pelo nó A *para o usuário* representado pelo nó B, o rótulo nas arestas indicam as *tags* utilizadas para a concessão de autoridade e o peso na

aresta representa quanto o usuário A considera o usuário B uma autoridade no assunto relacionado às *tags*.

Em termos do esquema de *ranking* utilizando autoridades, não se encontrou nenhuma referência na literatura a respeito, portanto, neste trabalho o cálculo do *score* de um documento foi realizado com base nas autoridades categorizadas com os termos da consulta, gerando um esquema denominado de *AuthorityRank* (AR). Para entender esse processo, pode-se considerar um exemplo de uma consulta com os termos `information` e `retrieval` (trata-se da conjunção dos dois termos, neste exemplo). Inicialmente, a abordagem de *ranking* restringe a cadeia de autoridades de forma que a sua topologia reflita somente as concessões de autoridade realizadas com *tags* pertencentes aos dois termos da consulta. A partir da cadeia de autoridades restringida pelas *tags* é realizado o cálculo do valor do *PageRank* à Priori (White e Smyth, 2003) de cada um dos usuários da cadeia, o qual serve como um peso para diferenciar os documentos categorizados por autoridade quando uma consulta contendo os termos `information` e `retrieval` é passada ao sistema. Uma vez calculado o valor do *PageRank* à Priori de cada autoridade, basta que esses valores sejam acrescidos ao *score* dos documentos, os quais podem ser calculados por meio de uma abordagem tradicional. É importante destacar que o valor do *PageRank* à Priori não está contido no modelo da cadeia de autoridades, mas é calculado em tempo de indexação a partir da topologia da cadeia de autoridades e dos pesos das arestas da cadeia.

Para avaliar a proposta do esquema de *ranking* AR, foi implementado um SRI denominado de *AuthoritySearch*, o qual utiliza a abordagem de *Folkauthority* para indicar a autoridade cognitiva das fontes de informação. As concessões de autoridade cognitiva no sistema foram simuladas com base em dados do SBF *Delicious*, cujos documentos são *bookmarks* disponibilizados pelos próprios usuários. O sistema *AuthoritySearch* permite a busca de documentos do sistema *Delicious*, no entanto o *ranking* das informações recuperadas é realizado com base na autoridade cognitiva dos usuários que disponibilizam tais informações, sendo utilizado o esquema de *ranking* AR para classificar os documentos em uma consulta. Esse esquema foi comparado com o tradicional esquema de *ranking* denominado de *tf-idf* (Baeza-Yates e Ribeiro-Neto, 1999; Manning et al., 2009), com a finalidade de aferir a diferença entre esses dois esquemas no que diz respeito à capacidade de apresentar documentos de relevância e qualidade nos primeiros resultados. Na metodologia adotada para a avaliação desses esquemas de *ranking*, foram elaborados cenários de busca nos quais definiu-se 9 consultas, distribuídas entre 7 usuários participantes, as quais foram passadas ao sistema *AuthoritySearch*, sendo obtidos resultados de busca para três esquemas de *ranking*.

Os critérios de relevância e de qualidade foram escolhidos para a comparação a partir de uma revisão sobre pesquisas que discutem quais critérios são utilizados pelos usuários para julgarem as informações recuperadas (Barry, 1994; Kargar, 2011; Knight e Burn, 2005; Lachica et al., 2008; Saracevic, 2007). Esses critérios foram escolhidos por representarem as dimensões da QI possíveis de serem aferidas pelos participantes nas condições em que as informações estavam representadas, pois os dados capturados do sistema *Delicious* continham somente a URL de cada documento e as *tags* utilizadas pelos usuários para categorizá-los. Dentre os critérios apontados pelos autores os quais não puderam ser aferidos pode-se citar a autoria, a representação, a compreensão e a objetividade.

Para avaliar quantitativamente os resultados dos experimentos, foi utilizada a métrica de NDCG (Järvelin e Kekäläinen, 2000), separando os dados em grupos de acordo com os seguintes atributos: (1) o critério para a avaliação, o qual é dividido em qualidade e relevância, (2) as consultas realizadas pelos usuários participantes, sendo que neste trabalho foram em número de nove, e (3) os esquemas de *ranking* utilizados para classificar os documentos apresentados, sendo que neste trabalho foram em número de três, dois dos quais utilizam um esquema baseado em autoridades e um que utiliza um esquema tradicional.

As médias para os valores de NDCG gerados na avaliação com os usuários participantes foram analisadas estatisticamente por meio do teste *T de Student*, com um nível de significância de 95% (Smucker et al., 2007). A partir desta análise, foi possível concluir que há uma diferença significativa com relação aos esquemas de *ranking* que consideram e que não consideram as autoridades cognitivas das fontes de informação, tanto para o critério de qualidade quanto para o critério de relevância, sendo que, na opinião dos usuários avaliadores, o esquema de *ranking* AR apresentou documentos contendo informação de maior relevância e qualidade nos primeiros resultados de busca. Além disso, foi possível concluir que as duas abordagens de *ranking* que utilizam as informações do *PageRank* à Priori da cadeia de autoridades propostas neste trabalho não produzem resultados com diferenças significativas com relação à relevância e à qualidade das informações apresentadas.

Este trabalho está organizado da seguinte maneira: no Capítulo 2, são apresentados os fundamentos teóricos para a implementação computacional de um SRI, mostrando as características dos modelos utilizados neste trabalho para a realização de tal tarefa. São também introduzidas as peculiaridades da RI no ambiente da *Web* e apresentados alguns algoritmos comumente utilizados para a Recuperação Social de Informação. Por fim, é discutido como a avaliação é realizada em SRIs e são apresentadas métricas para a realização de tal tarefa. No Capítulo 3, é apresentada a proposta de *Folkauthority* e

sua relação com o conceito de autoridade cognitiva, demonstrando como a utilização dessa abordagem pode ser benéfica para a RI. Além disso, são apontadas lacunas sobre a relação entre a abordagem de *Folkauthority* e a RI, mostrando questões que ainda necessitam de maior investigação. No Capítulo 4, é apresentado um modelo baseado em conjuntos para a definição de um sistema que utiliza a abordagem de *Folkauthority* e discutida a implementação da indexação e da busca nesse sistema. Além disso, também é apresentado o cálculo do *score* dos documentos utilizando a abordagem de *ranking AuthorityRank*. No Capítulo 5, são demonstrados os resultados obtidos com a avaliação de consultas ordenadas pelos esquemas de *ranking* que consideram e que não consideram o arcabouço estabelecido pela abordagem de *Folkauthority*, os quais demonstram que a consideração da abordagem de *Folkauthority* é uma proposta viável para a RI em SBFs. No Capítulo 6, são tecidas as conclusões do trabalho bem como apontadas suas limitações e alguns possíveis trabalhos futuros.

Capítulo

2

Recuperação de Informação

Até o início da década de 90, a maior parte das atividades relacionadas à Recuperação de Informação (RI) eram realizadas por pessoas diretamente relacionadas ao assunto, tais como bibliotecários, pesquisadores na área de ciência da informação e profissionais similares (Baeza-Yates e Ribeiro-Neto, 1999). Devido ao acesso mais restrito aos mecanismos automáticos de RI as pessoas preferiam ter acesso a uma informação por meio de outra pessoa ao invés de utilizar um Sistema de Recuperação de Informação (SRI) (Manning et al., 2009). A popularização do acesso à Internet e o crescimento da *Web* fomentaram o engajamento de diversas pessoas em atividades de RI como, por exemplo, ocorre na utilização de um mecanismo de busca na *Web*.

A disciplina de RI preocupa-se, principalmente, com a tarefa de encontrar itens de informação cujos conteúdos são de natureza não estruturada dentro de uma grande coleção de itens, de forma a satisfazer as necessidades de informação de um usuário. Esses itens são comumente documentos, sendo o termo *não estruturado* referente ao fato de que os dados não estão em uma forma diretamente compreensível por um computador. Assim sendo, uma questão primordial com relação à tarefa de RI é a de se estimar quais documentos são relevantes e quais não são, dada uma necessidade de informação. Nesse contexto, o termo *necessidade de informação* refere-se a um tópico ou assunto sobre o qual um usuário deseja saber mais, diferindo de uma *consulta*, a qual é um veículo de expressão da necessidade de informação de um usuário. Além disso, o termo *relevância* refere-se à percepção de que uma informação possui conteúdo que satisfaça (ou não) uma necessidade de informação pessoal (Fuhr, 2001; Manning et al., 2009).

Uma primeira questão relevante sobre a tarefa de RI é a forma de se expressar a necessidade de informação de um usuário. Em um cenário típico, na qual a tarefa de RI é comumente chamada de recuperação *ad hoc* (Baeza-Yates e Ribeiro-Neto, 1999; Manning et al., 2009), essa questão é tratada por meio da formulação de consultas pelo usuário, as quais são passadas ao SRI em uma determinada linguagem. Tais consultas especificam um conjunto de termos que denotam o significado da necessidade de informação do usuário (Salton et al., 1983). Apesar desse cenário típico, o usuário pode não possuir uma necessidade de informação bem formulada, de forma que possa, mesmo assim, estar realizando a tarefa de RI por meio da exploração de documentos em uma interface interativa. Costuma-se diferenciar essas duas situações dizendo que o usuário pode estar executando uma tarefa de recuperação (*retrieval*) ou uma tarefa de exploração/navegação (*browsing*). No entanto, considera-se que em ambos os casos esse usuário está envolvido em atividades relacionadas à área de recuperação de informação. Existem outras abordagens além dessa (Croft, 2000), no entanto a atenção será voltada a recuperação *ad hoc*.

A Figura 2.1 mostra uma interface de busca com suporte às duas situações acima citadas. Essa figura corresponde à primeira janela visualizada após a inicialização de um editor de documentos de texto, tendo como objetivo permitir ao usuário escolher um modelo de documento a ser editado. Considerando o usuário que acesse o sistema com o deliberado intuito de editar um documento no formato *carta*, a interface provê um campo para consulta textual, o qual é destacado na figura por meio de uma marca no formato de elipse. Da mesma forma, considerando o usuário que acesse o sistema sem uma intenção clara sobre qual formato de documento editar, a interface permite que este usuário explore os formatos pré-existentes, conforme destacado na figura por meio de uma marca no formato retangular. Desse modo, no primeiro caso tem-se uma noção bastante específica da necessidade de informação, sendo que uma consulta textual mais direta é adequada, enquanto no segundo caso a necessidade de informação não está tão bem definida, de forma que a interface permita uma exploração no espaço dos itens de informação (Jansen, 2010; Marchionini, 2006). Nesse sentido, (Jansen et al., 1998) cita que aumentando o número de recursos oferecidos pela linguagem, ou pela interface de consulta, para denotar uma necessidade de informação, ou, mesmo, aumentando o engajamento do usuário na elaboração de uma consulta, é possível conseguir melhores resultados na tarefa de RI.

Uma segunda questão sobre a tarefa de RI a ser destacada é a necessidade de processamento rápido de uma consulta. A forma tradicional de se evitar uma busca linear no conjunto de termos dos documentos é a construção de um *índice invertido* (Zobel e Moffat, 2006). Esse índice é construído de forma a relacionar cada termo com uma lista

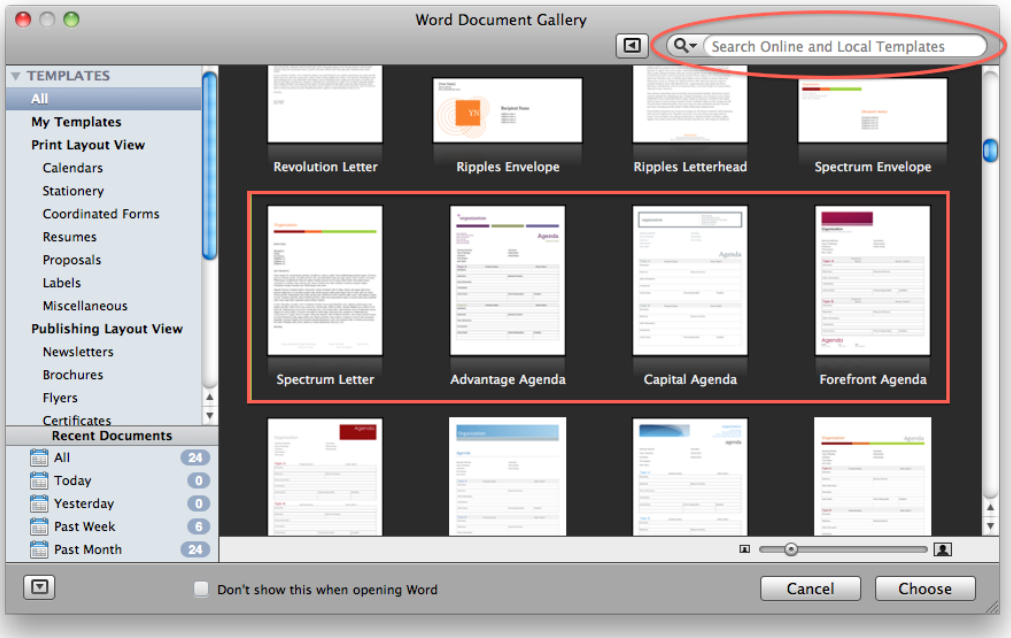

Figura 2.1: Diferenças na interface entre a recuperação e a exploração de informação.

de documentos nos quais o termo está presente. Os passos para a construção de um índice invertido incluem selecionar o conjunto de documentos a serem indexados, extrair uma lista de termos que estejam presentes em cada documento, realizar um pré-processamento linguístico de forma a obter uma lista de termos normalizados e, finalmente, indexar os documentos no índice invertido, relacionando cada termo com os documentos nos quais estejam presentes (Manning et al., 2009). Para exemplificar, a Figura 2.2 mostra o índice invertido para um conjunto de documentos contendo, dentre outros, os termos `document`, `indexing`, `information` e `retrieval`. Com o uso dessa estrutura de dados, pode-se conhecer os documentos que possuem determinado termo sem que seja necessária uma busca linear nos documentos.

Segundo Baeza-Yates e Ribeiro-Neto (1999), uma terceira questão central relacionada aos SRIs é a necessidade de se estimar quais documentos são relevantes e quais não são. Tal estimativa é dependente de um *algoritmo de ranking*, o qual realiza uma ordenação dos documentos. Nesse processo, é fundamental encontrar um peso para os termos em cada documento. O algoritmo de *ranking* calcula então um *score* para os documentos, tendo como base um parâmetro de similaridade entre os termos de cada documento e a consulta (Singhal, 2001). Os documentos são então ordenados com base nos seus *scores* e apresentados ao usuário, sendo que os documentos no topo da lista são os mais relevantes,

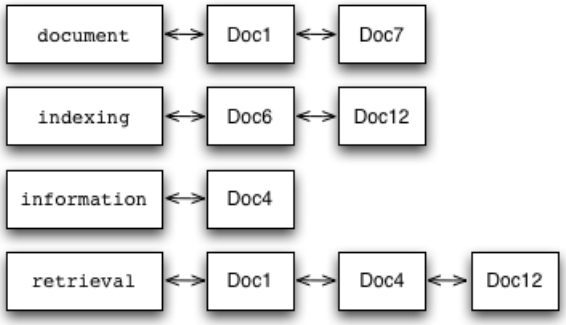

Figura 2.2: Índice invertido para um conjunto hipotético de documentos.

de acordo com o *esquema de ranking* empregado. A Figura 2.3 mostra uma interface de busca que explicita ao usuário o valor do cálculo da relevância de cada documento.

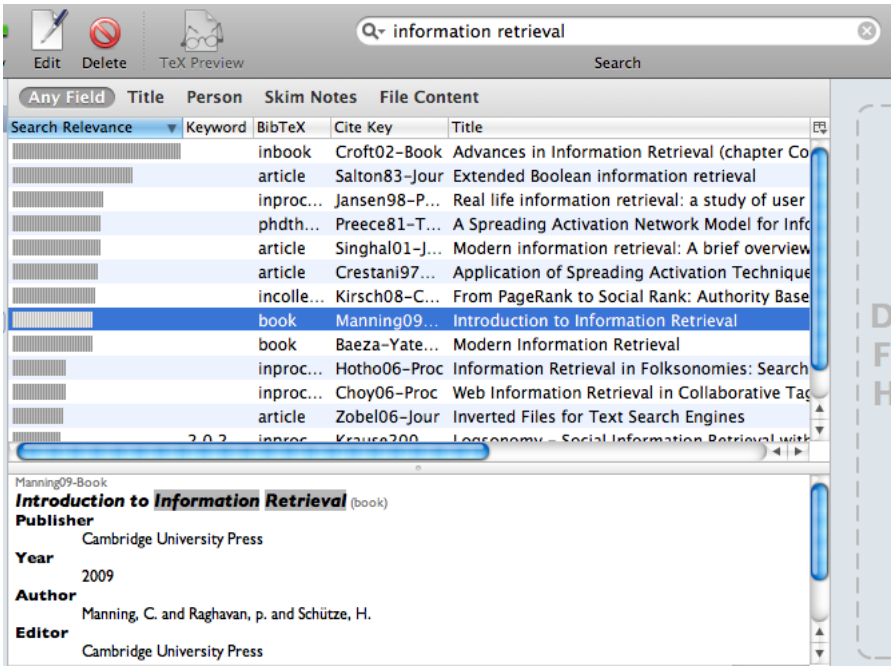

Figura 2.3: Resultados de busca com base em um *ranking*.

A pesquisa em RI passou a tratar e influenciar questões de modelagem (Singhal, 2001), classificação e categorização de documentos (Berkhin, 2002), indexação distribuída (Melnik et al., 2001), interface e interação com o usuário (Petratos, 2006), filtragem colaborativa e recomendação (Adomavicius e Tuzhilin, 2005), dentre outras. O desenvolvimento dos conceitos de *Web 2.0* e *Web* Social tem exigido um esforço no sentido de encontrar maneiras de melhorar a tarefa de RI explorando aspectos desses dois conceitos (Kim e

Kwon, 2009; Vlahovic, 2011; Zhou et al., 2008). A *Web 2.0* e a *Web* Social referem-se a dois novos paradigmas de *Web* mais interativos e colaborativos. Esses dois paradigmas são caracterizados principalmente pelo aprimoramento dos processos de trabalho coletivo e de produção e compartilhamento de informação (Primo, 2007). É nesse contexto que este trabalho foi desenvolvido e, portanto, é necessário apresentar as peculiaridades da área de RI na *Web* 2.0 e na *Web* Social. Deste modo, a seguir serão discutidos os modelos para RI, os processos de indexação, *ranking* e busca, a RI na *Web 2.0*, a questão da Recuperação Social de Informação (RSI) e, por fim, os métodos de avaliação de SRIs, uma vez que no desenvolvimento deste trabalho será necessário realizar a comparação entre diferentes abordagens para RI.

2.1 Modelos para Recuperação de Informação

Os modelos para RI especificam como os documentos e a necessidade de informação do usuário devem ser representados e comparados a fim de produzir um resultado de busca (Croft, 2000). Baeza-Yates e Ribeiro-Neto (1999) definem formalmente um modelo para RI, o qual é caracterizado na definição seguinte.

Definição Um modelo de Recuperação de Informação é uma 4-upla $\{D, Q, F, R(q_i, d_j)\}$ onde:

- D é um conjunto formado pelas representações dos documentos de uma coleção;
- Q é um conjunto formado pelas representações das necessidades de informação dos usuários;
- F é um arcabouço para a modelagem da representação dos documentos, das consultas e das suas relações;
- $R(q_i, d_j)$ é uma função de *ranking*, a qual associa um número real com uma consulta $q_i \in Q$ e uma representação do documento $d_j \in D$.

De acordo com Baeza-Yates e Ribeiro-Neto (1999), os três modelos clássicos da RI são o modelo booleano, o modelo de espaço vetorial e o modelo probabilístico. Esses modelos são diferenciáveis pelos elementos da 4-upla. Por exemplo, o modelo booleano utiliza a teoria de conjuntos para representar seus documentos (Frakes e Baeza-Yates, 1992), o modelo de espaço vetorial utiliza ferramentas da álgebra linear (Berry et al., 1999) e o modelo probabilístico utiliza a estimativa da probabilidade de um documento ser julgado como relevante (Fuhr, 1992).

Para Turtle e Croft (1992), a palavra *representação* é utilizada no sentido matemático ou lógico e refere-se a um conjunto de axiomas e regras de inferência que permitem a derivação de novas representações. Um modelo então restringe a maneira como a representação dos documentos interagem. Alguns autores incluem também outros modelos como sendo importantes, tais como os modelos de rede de inferência e os modelos de linguagem (Croft, 2000; Fuhr, 2001; Singhal, 2001).

A seguir serão apresentados os modelos booleano e de espaço vetorial, por serem utilizados no desenvolvimento desta pesquisa. O modelo booleano foi utilizado para selecionar os documentos que atendem a uma consulta, enquanto para o cálculo do *score* de um documento foi utilizado o modelo de espaço vetorial.

2.1.1 O Modelo Booleano

O modelo booleano para RI é um modelo no qual podemos representar uma consulta na forma de uma expressão booleana de termos, utilizando os operadores AND ($\wedge$), OR ($\vee$) e NOT ($\neg$). Para exemplificar o uso desse modelo, considere que o intuito é recuperar, dentre um conjunto de artigos sobre RI na língua inglesa, aqueles que possuem os termos `user` $\wedge$ `relevance` $\wedge \neg$ `quality`. Os termos contidos em cada documento desse exemplo estão relacionados na Tabela 4.2. Uma alternativa para resolver esse problema é realizar uma varredura linear no conteúdo dos documentos, verificando e marcando quais possuem os termos `user` e `relevance` e não possuem o termo `quality` (Manning et al., 2009). Uma abordagem que utilize um autômato para reconhecimento de padrões nos textos de cada documento leva tempo proporcional ao tamanho da soma dos termos encontrados em cada documento (Aho e Corasick, 1975). Uma forma de evitar tal varredura é utilizar uma matriz de incidência, a qual é ilustrada na Tabela 4.2.

Pode-se observar pela Tabela 2.2 que para responder a essa consulta é possível representar seus componentes `user`, `relevance` e `quality` respectivamente como os vetores $\vec{t}_{user} = (1, 1, 0, 1, 0, 1, ...)$, $\vec{t}_{relevance} = (0, 1, 0, 1, 1, 1, ...)$ e $\vec{t}_{quality} = (0, 1, 0, 1, 1, 0)$ (Manning et al., 2009). Esses vetores indicam a existência ou não do documento d_j no termo t_i e estão representados na Tabela 2.2 pelas linhas correspondentes aos termos. Dessa forma, a consulta `user` $\wedge$ `relevance` $\wedge \neg$ `quality` $= 110101(...) \wedge 010111(...) \wedge 101001(...) = 000001(...)$, o que significa que, considerando os seis primeiros documentos, somente o documento Doc6 atende à consulta.

A necessidade de se relacionar o conjunto de termos com os documentos nos quais os termos ocorrem é também justificada quando se considera a natureza não estruturada dos documentos. Segundo Frakes e Baeza-Yates (1992), diferentemente de dados armazenados

Documentos	Termos
Doc1	..., `retrieval`, `user`, ...
Doc2	..., `quality`, `relevance`, ...
Doc3	...
Doc4	..., `quality`, `relevance`, `retrieval`, `user`, ...
Doc5	..., `quality`, `relevance`, ...
Doc6	..., `relevance`, `retrieval`, `user`, ...
...	...

Tabela 2.1: Termos nos documentos

	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**	**Doc6**	...
...	...	...	...	...	...	...	...
`quality`	0	1	0	1	1	0	...
`relevance`	0	1	0	1	1	1	...
`retrieval`	1	0	0	1	0	1	...
`user`	1	0	0	1	0	1	...
...	...	...	...	...	...	...	...

Tabela 2.2: Matriz de incidência

em um banco de dados tradicional, os dados no contexto da RI geralmente não possuem um conteúdo uniforme. Nessas circunstâncias, um documento pode ser descrito por um conjunto de palavras chamadas de *termos de índice*. Esses termos de índice são armazenados de forma ordenada em uma estrutura de dados chamada de *índice invertido*. Nessa abordagem, mantém-se um conjunto de termos, chamado de dicionário, e uma lista de documentos associados com cada termo do dicionário. Por motivos de desempenho, ambas estruturas são mantidas ordenadas alfabeticamente e por *ID* do documento (Frakes e Baeza-Yates, 1992; Zobel e Moffat, 2006).

De acordo com Baeza-Yates e Ribeiro-Neto (1999), um termo de índice é uma palavra cujo significado auxilia a descrever o tema principal do documento. Esse mesmo autor destaca que nem todos os termos de índice de um documento possuem a mesma capacidade de descrever seu conteúdo, sendo que alguns termos são mais vagos do que outros. Dessa forma, considerando um conjunto de termos de índice T, o modelo de RI deve possuir um mecanismo para decidir sobre o peso $w_{i,j} \geq 0$ dos termos de índice t_i para descrever o conteúdo de cada documento d_j. Essa questão será discutida com maiores detalhes mais adiante. Por hora, é importante notar a seguinte definição (Baeza-Yates e Ribeiro-Neto, 1999).

Definição Seja k o número de termos e n o número de documentos em um sistema. Seja também t_i um termo de índice qualquer. $T = \{t_1, t_2, ..., t_k\}$ é o conjunto de todos os termos de índice e $D = \{d_1, d_2, ..., d_n\}$ o conjunto e todos os documentos. Um peso $w_{i,j} > 0$ é associado a cada termo de índice $t_i \in T$ que esteja contido em um documento $d_j \in D$. Para um índice de termo que não apareça em um documento, $w_{i,j} = 0$. Ao documento d_j é associado um vetor de pesos $\vec{d_j}$ representado por $\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{k,j})$, o qual contém um peso para cada um dos termos contidos no documento d_j. Além disso, considera-se a função g_i tal que $g_i(\vec{d_j}) = w_{i,j}$. É acrescida também a definição do vetor $\vec{t_i} = (w_{i,1}, w_{i,2}, ..., w_{i,n})$, o qual contém os pesos dos documentos que contém o termo t_i.

Sintetizando as definições de Baeza-Yates e Ribeiro-Neto (1999), Turtle e Croft (1992) e Fuhr (2001), podemos afirmar que o modelo booleano é aquele em que $w_{i,j} \in \{0, 1\}$ assume valores binários, dependendo da existência ou não do termo t_i no documento d_j. O fato do termo t_i estar presente no documento d_j é representado no modelo por $w_{i,j} > 0 = 1$. Baeza-Yates e Ribeiro-Neto (1999) e Manning et al. (2009) discutem o processamento de consultas nesse modelo utilizando um índice invertido. A utilização do algoritmo de processamento de consultas apresentado por Manning et al. (2009) pode ser exemplificada com a consulta `relevance` $\wedge$ `quality`. O processamento dessa consulta, mostrado em sua forma conjuntiva, é realizado seguindo os seguintes passos:

1. Monte uma lista com os documentos que possuem o termo $t_{relevance}$, ordenada por ID do documento;

2. Monte uma lista com os documentos que possuem o termo $t_{quality}$, ordenada por ID do documento;

3. Faça a intersecção das duas listas conforme o Algoritmo 1.

Na listagem do Algoritmo 1 são mantidos dois ponteiros que percorrem os elementos das listas, simultaneamente, em tempo linear proporcional ao número total de elementos das listas. Em cada passo, é comparado o ID dos documentos referenciados pelos dois ponteiros. No caso das referências serem as mesmas, o ID do documento é adicionado a lista *intersec*, caso contrário o ponteiro referenciando o documento com o menor ID é avançado. No caso das listas possuírem tamanhos x e y, a execução do algoritmo realiza $O(x + y)$ operações (Manning et al., 2009). Na prática, a complexidade da consulta é $\Theta(n)$, onde n é a quantidade de documentos. Segundo Manning et al. (2009) o ganho com relação a uma varredura linear é somente em termos de uma constante (bastante grande na prática), no entanto isto não altera a complexidade assintótica de tempo Θ. Dessa forma, existem pesquisas em uma área denominada de *otimização de consulta* que

Algoritmo 1. Intersecção de duas listas de documentos

Entrada: p_1 e p_2, ponteiros para os elementos das listas
Saída: Lista com os ID's dos documentos resultantes da intersecção

```
intersec ← ∅
enquanto p1 ≠ NULL e p2 ≠ NULL faça
    se docID(p1) = docID(p2) então
        Adiciona(intersec, docID(p1))
        p1 ← proximo(p1)
        p2 ← proximo(p2)
    fim
    senão se docID(p1) < docID(p2) então
        p1 ← proximo(p1)
    fim
    senão
        p2 ← proximo(p2)
    fim
fim
retorna intersec
```

visam otimizar a ordem em que a consulta é avaliada, consequentemente, melhorando esse tempo computacional. No entanto esse assunto está além do escopo deste trabalho.

O modelo booleano possui diversas limitações, dentre as quais destaca-se:

- A impossibilidade de realizar um *ranking* dos documentos recuperados, uma vez que a única informação sobre o peso $w_{i,j}$ de um termo em um documento é um valor binário (Manning et al., 2009).
- A dificuldade em prever e controlar o número de documentos recuperados em uma consulta (Salton et al., 1983), na qual o número de documentos recuperados pode ser muito grande ou muito pequeno (Baeza-Yates e Ribeiro-Neto, 1999);
- Algumas estruturas linguísticas são representadas de forma inábil no modelo (por exemplo, variações morfológicas de um termo) ou não podem ser totalmente representadas (frases, por exemplo) (Turtle e Croft, 1992).

No entanto, essas limitações não impediram que o modelo fosse largamente utilizado em aplicações reais. Diversas técnicas, que estão além do escopo deste trabalho, são utilizadas para contornar essas limitações (Turtle e Croft, 1992). Dentre as vantagens atribuídas ao modelo booleano destacam-se o formalismo por trás do modelo, sua simplicidade e o alto padrão de desempenho que pode ser mantido com a sua utilização (Baeza-Yates e Ribeiro-Neto, 1999; Fuhr, 2001; Salton et al., 1983).

2.1.2 O Modelo de Espaço Vetorial

Conforme explicam Baeza-Yates e Ribeiro-Neto (1999), o uso de pesos binários associados aos termos de um documento é considerado limitado. Assim, o modelo de espaço vetorial (também chamado de modelo vetorial) apresenta um arcabouço no qual uma relação parcial entre um termo e um documento pode ser medida. Neste modelo os documentos e as consultas são representados como vetores em um espaço k-dimensional, no qual cada dimensão corresponde a um termo[1]. Cada elemento do vetor de pesos $\vec{d_j}$ corresponde à similaridade do documento d_j com os termos pertencentes à consulta (Turtle e Croft, 1992), refletindo a importância de cada termo para a semântica do documento (Berry et al., 1999). O principal resultado dessa abordagem é a possibilidade de se utilizar um algoritmo de *ranking*, com a finalidade de ordenar os documentos por similaridade com relação aos termos de uma consulta. Com essa característica, pode-se conseguir uma melhor satisfação quanto à necessidade de informação de um usuário. A fim de formalizar o modelo de espaço vetorial, será utilizada a definição baseada nos autores Baeza-Yates e Ribeiro-Neto (1999), descrita a seguir:

Definição No modelo de espaço vetorial o peso $w_{i,j}$ associado a um termo de índice t_i presente no documento d_j é um valor real positivo. Além disso, os termos de uma consulta também possuem pesos, sendo $w_{i,q} \geq 0$ o peso associado com o termo t_i da consulta q. Dessa forma, a consulta é definida por um vetor $\vec{q} = (w_{1,q}, w_{2,q}, ..., w_{k,q})$ no qual k é o número de termos de índice e $g_i(\vec{q}) = w_{i,q}$.

Considerando um vetor de pesos $\vec{d_j} = (w_{1,j}, w_{2,j}, ..., w_{k,j})$ e um vetor consulta $\vec{q}$, a similaridade entre esses dois vetores no modelo de espaço vetorial é comumente estimada por meio da medida de cosseno entre os vetores consulta $\vec{q}$ e de pesos $\vec{d_j}$, sendo possível um cálculo de similaridade entre a consulta e cada documento (Keen, 1971). No entanto, outras medidas de similaridade também podem ser utilizadas (Berry et al., 1999; Robertson e Jones, 1976). A medida de cosseno é definida pela Equação 2.1.

$$sim(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} \tag{2.1}$$

Nessa equação, o numerador representa o produto interno dos vetores $\vec{d_j}$ e $\vec{q}$, definido por $\sum_{j=1}^{k} w_{k,j} \times w_{k,q}$ e o denominador representa o produto entre as normas euclidianas de $\vec{d_j}$ e de $\vec{q}$. A norma euclidiana de um vetor $\vec{v}$ de k componentes é definida por $\sqrt{\sum_{j=1}^{k} \vec{v_i}^2}$.

[1] De acordo com Turtle e Croft (1992), ao realizar a indexação dos termos de índice é possível também armazenar o campo do texto onde o termo aparece. Por exemplo, um mesmo termo `relevance` que apareça no título e no corpo de um documento seria indexado com os termos de índice $t_{relevance.titulo}$ e $t_{relevance.corpo}$. Nesse caso, cada termo de índice corresponde a uma dimensão dos vetores.

O efeito do denominador nessa equação é o de realizar uma normalização nos vetores $\vec{d_j}$ e $\vec{q}$ para os seus respectivos vetores unitários $\hat{d_j}$ e $\hat{q_j}$, de forma que a Equação 2.1 pode ser reescrita pela Equação 2.2

$$sim(\vec{d_j}, \vec{q}) \ = \ \hat{d_j} \cdot \hat{q} \tag{2.2}$$

Essa definição levanta a questão de como definir o valor de cada dimensão do documento, isto é, os valores de $w_{k,j}$. Conforme descrito anteriormente, um mesmo termo de índice pode ser considerado melhor ou pior para descrever o significado de diferentes documentos. Dessa forma, não é razoável atribuir o mesmo valor ao peso associado aos termos em um documento. Além disso, considerando uma consulta qualquer, um documento (ou parte de um documento) que contenha muitos termos presentes na consulta tem mais relação com essa do que um documento que contenha poucos termos. Para representar essa situação, no modelo de espaço vetorial é comumente utilizado um peso $w_{k,j}$, proporcional à frequência do termo t_k no documento d_j. Isto é, o peso associado a cada termo presente em um documento depende do número de ocorrências do termo no documento (Manning et al., 2009). A abordagem mais simples é utilizar, como valor para o peso, a própria frequência do termo no documento (*term frequency*), denotado por $tf_{i,j}$, na qual os índices denotam, respectivamente, o termo e o documento.

Com base nessa ideia, é também observado que os termos que aparecem com muita frequência na coleção de documentos são considerados menos importantes do que os termos que aparecem com menos frequência, no sentido de proverem alguma informação sobre o conteúdo do documento. Por exemplo, em uma coleção de artigos na língua inglesa sobre recuperação de informação é provável que os termos `information` e `retrieval` apareçam em quase todos os documentos, o que implica que sejam termos com pouca capacidade de diferenciar os documentos da coleção. Com base nisso, é introduzido um fator de atenuação no peso, no qual termos muito frequentes na coleção de documentos recebem um desconto maior do que os termos cuja frequência seja menor. No cálculo desse fator de atenuação utiliza-se um valor denotado por df_i, que representa o número de documentos na coleção que contém o termo t_i. Com base nesse valor, calcula-se o fator chamado *inverse document frequency*, demonstrado na Equação 2.3:

$$idf_i = \log \frac{n}{df_i} \tag{2.3}$$

Nota-se que o valor de idf_i para um termo menos frequente é maior do que para um termo mais frequente. Utilizando as definições de $tf_{i,j}$ e de idf_i define-se um esquema de cálculo dos pesos chamado de *tf-idf*, no qual o peso $w_{i,j}$ é dado pela Equação 2.4.

$$w_{i,j} = tf\text{-}idf_{i,j} = tf_{i,j} \times idf_i \tag{2.4}$$

Dadas essas definições, pode-se considerar os documentos como um vetor com cada componente correspondendo a um termo e cujo valor é o peso do termo no documento, conforme definido na Equação 2.4. Caso um determinado termo não apareça em um documento seu valor é 0. Dessa forma, conforme exposto por Manning et al. (2009), é possível definir um algoritmo de *ranking* utilizando a Equação 2.1, o qual é mostrado na listagem Algoritmo 2.

Algoritmo 2. Algoritmo básico para busca no espaço vetorial

Entrada: Consulta q
Saída: m primeiros componentes do ranking

```
float Scores[n] = 0
Inicialize o vetor de fatores de normalização Normas[n]
para cada termo t_i em q⃗ faça
    Calcule w_{i,q}
    Recupere os pesos de cada documento d_j que contém o termo t_i
    para cada par (d_j, w_{i,j}) faça
        Scores[d_j] += w_{i,j} × w_{i,q}
    fim
fim
para cada documento d_j em D faça
    Scores[d_j] = Scores[d_j]/Normas[d_j]
fim
retorna m primeiros componentes de Scores[]
```

Na listagem Algoritmo 2, o *array* $Normas[\,]$ é inicializado com os fatores de normalização, enquanto o *array* $Scores[\,]$ armazena os *scores* de cada documento. O *loop* mais externo itera sobre cada termo t_i da consulta, atualizando o valor de $Scores[\,]$ para cada documento que possua esse termo. O *loop* interno itera sobre os pesos associados a cada documento d_j que contenha o termo t_i, somando à $Score[j]$ a contribuição de cada um desses pesos, os quais podem ser o valor de $tf_{i,j}$ ou o valor de $tf\text{-}idf_{i,j}$.

Nessa abordagem, faz-se necessário o armazenamento desses valores juntos a lista de *postings* de cada termo, a qual é definida como a lista de documentos que esse termo aparece (essa lista é representada pelo vetor de pesos $\vec{t_i}$, definido anteriormente). Assim que terminado o cálculo dos *scores* dos documentos, basta retornar os m documentos com maior *score* presentes na lista, o que requer o uso de uma estrutura de dados de fila de prioridade, a qual pode ser implementada como um *heap*. Esse *heap* é construído utilizando não mais do que $2n$ comparações e permite que um elemento seja encontrado

utilizando não mais do que $O(\log n)$ comparações (Manning et al., 2009). Cabe ainda a observação do valor de $w_{i,q}$, o qual permite distinguir um peso (*boost*) para cada termo da consulta (Berry et al., 1999). O termo *boost* é utilizado neste trabalho para indicar um valor que diferencia o *score* de alguns documentos com base em determinados termos de uma consulta. Por exemplo, o termo `relevance` contido no documento Doc6 (Tabela 4.2) pode possuir um *boost* com valor positivo, enquanto os outros termos possuam um *boost* igual à zero. Nesse caso, o valor positivo de *boost* pode ser multiplicado ao valor de *score* do documento Doc6 para uma consulta contendo o termo `information`, considerando que os valores de *boost* e *score* estejam normalizados na mesma escala.

Segundo Baeza-Yates e Ribeiro-Neto (1999), as principais características vantajosas no modelo de espaço vetorial são as seguintes:

- O esquema de peso dos termos, o qual melhora o desempenho da recuperação de informação no modelo;
- Sua capacidade de medir uma relação parcial entre um termo e um documento, a qual permite a recuperação de documentos que se aproximam das condições impostas pela consulta;
- A fórmula de *ranking* por cosseno, a qual permite ordenar os documentos de acordo com o grau de similaridade com a consulta.

O modelo de espaço vetorial é bastante utilizado nos dias atuais, principalmente em aplicações comerciais e na *Web*, dada sua simplicidade e seu desempenho (Melnik et al., 2001). O modelo fornece a capacidade de se obter resultados ordenados por relevância, os quais são difíceis de serem melhorados sem a utilização de técnicas de expansão de consulta ou *relevance feedback* (Baeza-Yates e Ribeiro-Neto, 1999). Muitos outros métodos de *ranking* foram comparados com o modelo de espaço vetorial, no entanto há um consenso de que o modelo de espaço vetorial é superior ou quase tão bom quanto as outras alternativas.

2.2 Recuperação de Informação na Web 2.0

O acesso à *Web* vem crescendo de forma acelerada nos últimos anos e, em especial, a utilização de SRIs nesse ambiente tornou-se extremamente popular (Kargar, 2011). Manning et al. (2009) apontam que a característica essencial que leva aos desafios encontrados por SRIs na *Web* está relacionada a enorme quantidade de informação publicada sem controle de autoria. Além disso, essa grande quantidade de informação está publicada

de forma bastante heterogênea no que diz respeito à sua forma (língua escrita e formato do documento, por exemplo) e conteúdo (assunto tratado no documento, por exemplo). Essa heterogeneidade demanda dos SRIs para *Web* – também chamados de mecanismos de busca na *Web* – novas formas de operações sobre as informações, tais como o uso de algoritmos de *stemming* para várias línguas.

Outra preocupação com relação à elevada taxa de publicação de informação na *Web* refere-se à qualidade e à confiança dessas informações (Kargar, 2011). Devida a falta de controle sobre a autoria e o conteúdo das informações publicadas, na *Web* é possível encontrar verdades, mentiras, contradições, suposições e afirmações em uma grande escala (Manning et al., 2009). Questões como *confiança* (*trust*), popularidade, autoridade e reputação são bastante evidentes na *Web* (Golbeck e Hendler, 2006; Kazai e Milic-Frayling, 2008), pois é difícil para o usuário de um SRI na *Web* julgar essas questões com relação aos documentos, dada a falta de controle de qualidade e autoria. Além do mais, tamanha é a quantidade de informação publicada que a *Web* hoje está sujeita a um fenômeno chamado de "sobrecarga de informação" (Himma, 2007; Vlahovic, 2011), o qual está relacionado à incapacidade humana de absorver, interpretar e gerenciar tamanha quantidade de informação de maneira apropriada e produtiva.

Com relação aos aspectos técnicos da RI no ambiente da *Web*, pode-se destacar a necessidade extra da operação de *crawling* para a coleta e o processamento dos documentos. *Crawling* é o processo pelo qual os dados são coletados da *Web*, a fim de que sejam processados e indexados no mecanismo de busca (Bailey et al., 2003; Castillo, 2005; Cothey, 2004). Segundo Manning et al. (2009), o objetivo da técnica de *crawling* é coletar com rapidez e eficiência o máximo de páginas *Web* úteis, acompanhadas de suas respectivas referências de *hiperlink*. Por "páginas úteis" entende-se aquelas páginas com potencial de satisfazer alguma necessidade de informação dos usuários. Algumas características são obrigatórias em um sistema de *crawling*, as quais incluem a capacidade de lidar com estruturas de *links* maliciosos (*spams*) e de respeitar as políticas de indexação de recursos impostas pelos meta-dados dos documentos coletados, dentre outras características (Hirai et al., 2000). Outras características são desejadas dos sistemas de *crawling*, tais como robustês, qualidade, eficiência, escalabilidade e recência (Baeza-yates e Castillo, 2002). A questão é que a natureza distribuída, a organização em hipertexto e a falta de controle dos documentos presentes na *Web* oferecem alguns desafios para a implementação efetiva de um sistema de *crawling*, o qual de fato atinja os objetivos estabelecidos. No entanto, uma discussão mais aprofundada desse assunto está além do escopo deste trabalho.

2.2.1 Folksonomias e a RI

Os SBFs se aproveitam da característica social da *Web 2.0* para aprimorar a organização, o gerenciamento e a RI nesse ambiente (Trant, 2009). Os aspectos discutidos por pesquisadores como benefícios inerentes às Folksonomias, e que se relacionam com a RI, são: i) o senso de comunidade gerado pela utilização da técnica – o qual ocorre em poucos (ou nenhum outro) esquema de gerenciamento/recuperação de informação, ii) a opinião explícita de um conjunto de usuários a respeito dos conteúdos disponibilizados, ao invés de um visão única ou centralizada, relacionada ao texto do documento ou geradas por classificações automáticas/de especialistas – característica a qual parece favorecer o serendipismo nesses sistemas e, iii) a possibilidade de refletir, quase que em tempo real, mudanças no vocabulário utilizado para se expressar a respeito dos recursos (Golder e Huberman, 2006; Halpin et al., 2007; Sen et al., 2006).

Segundo Zhou et al. (2008) e Pereira e da Silva (2008b), a utilização da técnica de Folksonomia pode esbarrar em alguns problemas. Os autores afirmam que pelo fato das *tags* serem geradas pelos próprios usuários (e não baseado no conteúdo do documento, como geralmente acontece com os termos de índice) a diferença entre o conhecimento dos usuários deveria ser considerada no momento da categorização. Alguns usuários podem se diferenciar em seu conhecimento sobre determinado assunto e, consequentemente, isso afeta a qualidade de suas categorizações sobre os documentos (Rosch, 1978). Nesse mesmo contexto, este trabalho explora uma abordagem na qual o conhecimento do usuário do SBF sobre determinado assunto é levado em consideração no momento da RI.

Muitos trabalhos discutem a questão da RI utilizando a técnica de Folksonomia desde que o termo foi cunhado em 2004 (Vander Wal, 2007) e dentre esses trabalho alguns mais recentes podem ser destacados. Por exemplo, Zhou et al. (2008) propõem a utilização do modelo probabilístico para a RI em SBFs, utilizando um modelo generativo baseado no domínio de interesse dos usuários para descrever o SBF, além de também categorizar os usuários baseados nesses domínios. Schenkel et al. (2008) apresentam uma revisão sobre trabalhos na área de *tagging*/Folksonomia e discutem algumas dimensões sobre a técnica de Folksonomia que podem ser levados em consideração ao se projetar um SBF. Os autores enfatizam que a "indexação manual", a qual é realizada ao se utilizar a técnica de Folksonomia, apesar de bastante popular não representa de fato uma inovação. No entanto, eles apontam como vantagem da utilização dessa técnica a facilidade e a simplicidade de sua utilização, fazendo com que – nas palavras dos autores – "um esquema tradicional de organização de informação pareça uma técnica da idade da pedra, efetivo porém muito desconfortável". Trant (2009) apresenta a primeira vasta revisão bibliográfica

tratando exclusivamente sobre o assunto de Folksonomia. Em seu trabalho ele identificou, diferenciou e definiu três abordagens baseadas na referida técnica, denominadas de *folksonomia* – "*relacionada às tags e à indexação e recuperação de informação*", de *tagging* – "*relacionada ao comportamento dos usuários*" e de *SBFs* – "*relacionada ao arcabouço sócio-técnico*" (Trant, 2009). Chen e Zhang (2009) utilizaram dados de dois sistemas reais para avaliar a RI utilizando a técnica de *Folksonomia* e reportaram melhorias sobre a métrica NDCG na ordem de 10% com as abordagens que utilizaram *tags*. Eklund et al. (2010) reportam um SBF para um museu digital cuja principal característica é permitir a exploração e a recuperação de informação por meio de uma interface que utiliza uma técnica de síntese formal de conceitos baseado nos atributos dos documentos. Kim (2010) apresenta uma abordagem para a recuperação de texto e imagem baseada na técnica de *tagging*, na qual o uso das *tags* é realizado durante todo o processo de busca. O autor indica resultados sugerindo que as *tags* são mais utilizadas para reformulação/expansão das consultas e dessa forma conclui que manter as *tags* visíveis por todo o processo de busca pode auxiliar a RI nesses sistemas. Maniu et al. (2011) afirmam que a consideração das relações sociais em um SBF pode auxiliar a gerar melhores resultados de busca e apresentam uma abordagem com potenciais de escalabilidade na qual um modelo das relações sociais dos usuários é utilizado. Yong (2011) propõe uma abordagem baseada no modelo de linguagem para a RI, na qual utiliza informação sobre a rede social formada pela atividade de categorização dos usuários para melhorar a RI em SBFs. Vlahovic (2011) discute os impactos da utilização da técnica de Folksonomia para a RI e conclui que a principal vantagem na utilização dessa técnica é a possibilidade de melhorar a precisão de resultados de busca, uma vez que as informações são indexadas por metadados gerados por meio da "inteligência coletiva dos usuários".

Como é possível perceber nesta revisão de literatura sobre a RI utilizando a técnica de Folksonomia, há uma certa unanimidade com relação à crença de que o conhecimento coletivo gerado pela atividade de categorização por parte dos usuários, bem como as relações sociais que são formadas a partir dessa atividade, podem ser utilizadas para beneficiar a tarefa de RI em SBFs. Sistemas que utilizam dados das relações sociais do Sistema de Recuperação de Informação são chamados de Sistemas Sociais de Recuperação de Informação (SSRIs) (Cogo e da Silva, 2010). Além disso, existe uma área denominada de Recuperação Social de Informação (RSI) que apresenta modelos teóricos e ferramentas para se projetar e avaliar SSRIs, a qual será apresentada na subseção seguinte.

2.2.2 Recuperação Social de Informação

A ideia de se encontrar informações utilizando as relações sociais é relativamente antiga. Antes da ascensão dos meios de comunicação modernos o único método possível para se obter alguma informação era perguntando a alguém – um conhecedor do assunto, um amigo ou até mesmo um bibliotecário. Manter-se informado dependia das relações sociais mantidas pelas pessoas. A ideia de autoridade está também relacionada com essa questão: não é apenas importante conhecer muitas pessoas, mas é mais importante ainda conhecer as pessoas certas, no sentido de ser possível realizar as perguntas corretas para se obter as respostas corretas (Kirsch et al., 2008). Antes da comunicação moderna essa era a maneira mais direta de se obter informação confiável e de qualidade. Com o advento da comunicação moderna, métodos e técnicas para a RI foram propostos. Estas primeiras técnicas eram baseadas, principalmente, no conteúdo dos documentos a serem recuperados. Com a introdução do hipertexto foi possível melhorar ainda mais a precisão dos resultados das buscas por meio de algoritmos que exploram metadados presentes nos documentos e a estrutura relacional do hipertexto, da qual é possível extrair uma noção numérica da popularidade (*authority*) de cada documento na rede formada dessa estrutura (Golovchinsky, 1997; Kleinberg, 1999).

Os modelos tradicionais de RI têm seu foco nos documentos, nas consultas e nas relações entre esses dois conceitos. Já os modelos de análise de redes sociais têm seu foco em indivíduos e suas relações: famílias, amigos, conhecidos, colaboradores ou parceiros sexuais. SRIs tradicionais não incluem em seus modelos conceituais tais indivíduos, nem os papéis destes indivíduos como usuários do sistema ou como autores dos documentos recuperados. Da mesma forma, os modelos de redes sociais não incorporam os conceitos de necessidade de informação, consulta, termo de índice ou documentos passíveis de serem recuperados. Algumas pesquisas recentes demonstram que, apesar dos trabalhos tradicionais em RI na *Web* tratarem a tarefa de RI como uma atividade solitária (Evans e Chi, 2008), as interações sociais que ocorrem entre os usuários de um SRI cumprem um papel importante no processo de busca. Dessa forma, é possível notar um interesse sobre pesquisas que relacionem a RI com conhecimentos do domínio de Análise de Redes Sociais. No entanto, as pesquisas que chamam atenção para esse tema vêm de mais longa data, por exemplo, Karamuftuoglu (1998) argumenta que um problema fundamental com relação à RI é a necessidade de produção e consumo de conhecimento, defendendo o ponto de vista de que a produção de conhecimento é um trabalho fundamentalmente colaborativo e dependente das práticas de uma comunidade de participantes. O artigo aponta referências teóricas e reflexões para o projeto de SRIs que considera o esforço

coletivo de vários usuários para a concretização da tarefa de RI. Nessa mesma época, uma proposta de SRI foi apresentada, a qual discutia uma abordagem baseada na classificação coletiva de domínios específicos de informações contida em um sistema na *Web*, na qual era possível realizar a "recomendação" de recursos entre os usuários do sistema com interesses similares. O resultado do esforço para prover modelos de RI com informações sociais tem sido chamado de *Social Information Retrieval* – Recuperação Social de Informação (RSI)[2] – o qual refere-se a abordagens e tecnologias que dão suporte à colaboração e à socialização no processo de busca (Krause et al., 2008).

Nos anos de 2007 e 2008 duas coletâneas (Chevalier et al., 2008; Goh et al., 2007) foram publicadas contendo artigos cujo tema referenciavam exclusivamente o domínio da RSI. Nessas coletâneas, é possível encontrar o artigo de Kirsch et al. (2008), que propõem um modelo conceitual para a construção de um Sistema Social de Recuperação de Informação (SSRI), o qual contém três conceitos fundamentais: os indivíduos (pertencentes a estrutura social subjacente ao SRI); os recursos (isto é, os documento a serem recuperados); e a consulta. Nesse modelo, os indivíduos podem ser autores das informações contidas no SRI ou usuários dessas informações. A relação social que pode ocorrer entre os indivíduos no SRI é levada em consideração no momento da busca. Por exemplo quando dois ou mais indivíduos categorizam o mesmo recurso em um SBF é possível personalizar a busca desses indivíduos, bastando recalcular o *score* dos documentos apresentados na busca com base nessa relação. A necessidade de informação desses indivíduos pode ser expressa por uma consulta passada como entrada ao sistema. O sistema, por sua vez, pode automaticamente estimar a relevância dos documentos para este indivíduo baseado no conteúdo do documento e nas informações sociais relativas aos indivíduos. Desta forma, o domínio dos SSRIs incorpora tanto conceitos relacionados às redes e relações sociais quanto conceitos relacionados à RI. Trabalhos mais recentes também incluem (Foley e Smeaton, 2010), que apresenta uma discussão sobre dois conceitos relacionados ao tema de RSI – em especial sobre técnicas síncronas de RSI – os quais orientam o projeto e a construção de SSRI efetivos. Chi (2009) discute um modelo canônico de três estágios o qual representa um processo de busca em um SRI, destacando dois tipos de SSRIs: os sistemas de resposta (*social answering systems*), os quais utilizam a opinião de usuários com conhecimento em determinado assunto a fim de responder perguntas em domínios particulares; e os sistemas de *feedback*, os quais utilizam a popularidade e o *feedback* dos usuários sobre as informações para o cálculo do *ranking*. Kazai e Milic-Frayling (2008)

[2]O termo Recuperação Social de Informação é um termo que denota conceitos relacionados aos temas de *Colaborative Information Retrieval* – Recuperação Colaborativa de Informação e *Social Search* – Busca Social. A relação entre esses termos reside no fato de representarem abordagens para a recuperação/recomendação de informação que utilizam informações sociais entre os usuários do SRI.

apresentam uma abordagem na qual os modelos tradicionais de RI são estendidos para levar em conta a reputação e a confiança entre os usuários de um SRI.

Pode-se destacar também algumas abordagens apresentadas na literatura que apontam para a RSI. Por exemplo, os algoritmos *PageRank* (Page et al., 1998; White e Smyth, 2003) e HITS (Kleinberg, 1999; Li et al., 2002) são capazes de predizer um valor de autoridade em um grafo direcionado, contribuindo para a RI em ambientes de *hiperlink*. O modelo associativo para RI (Preece, 1981) baseia-se em grafos para representar a relação entre termos, autores e documentos em um SRI (Kirsch et al., 2008). Esses grafos são denominados de redes associativas e podem ser ponderados a fim de representarem a relação entre os termos, autores e documentos. Além disso, pode-se destacar a técnica de *spreading activation* para o cálculo do *score* dos documentos a serem recuperados em uma rede associativa (Crestani, 1997).

Com relação às abordagens relacionadas à RSI, o cálculo do algoritmo de *PageRank* à Priori (White e Smyth, 2003) para um grafo orientado merece destaque. Esse algoritmo foi utilizado neste trabalho para o cálculo da importância de uma autoridade em determinada *tag* na cadeia de autoridades. A Equação 2.5 mostra a fórmula para o cálculo do *PageRank* à Priori, na qual é definido um vetor de probabilidades à priori $p_{\mathrm{R}} = \{p_1, ..., p_{|V|}\}$ tal que a soma das probabilidades é 1. Cada dimensão p_v representa a importância relativa (ou probabilidade à priori) de se atingir o nó v a partir de um caminho aleatório no grafo. Neste trabalho $p_v = 1/x_{v,t}$, sendo $x_{v,t}$ o peso concedido à autoridade v na *tag* t. Assim sendo, a probabilidade de um nó v ser visitado *através* do nó u é $p(v|u)$.

Na proposta de esquema de *ranking* apresentada neste trabalho $p(v|u)$ é calculado com base no peso concedido pelo usuário u para a autoridade v (de fato, esse valor é a normalização da soma dos pesos que todos os usuários atribuíram à autoridade v, os quais são denotados por $d_{in}(v)$). O parâmetro β informa a probabilidade de, a cada iteração i do algoritmo, o caminho no grafo ser reiniciado desde o primeiro nó visitado.

$$pr_v^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v|u) \times pr_u^{(i)}\right) + \beta p_v \tag{2.5}$$

Neste trabalho será desenvolvida uma abordagem para a RI baseada no conceito de *Folkauthority*, que descreve um arcabouço onde os usuários do SRI indicam por meio de *tags* as competências e as habilidades cognitivas (isto é, a *autoridade cognitiva*) de cada indivíduo. Essas indicações explícitas sobre o conhecimento de cada indivíduo são levadas em consideração no momento do cálculo do *score* dos resultados de busca. Dessa

forma, para o desenvolvimento dessa abordagem buscou-se o respaldo teórico dos modelos clássicos de RI bem como daqueles que apontassem para o domínio da RSI.

2.3 Avaliação de Sistemas de Recuperação de Informação

Em SRIs, há diferentes maneiras de se processar os documentos antes da indexação, indexar os documentos, representar os documentos no índice e calcular a importância dos documentos para uma consulta. Dessa forma, é necessário também poder avaliar qual dessas abordagens (e também a forma como utilizá-las) resulta em maior eficiência para um SRI (Buckley e Voorhees, 2004). Conforme comentado no final da seção anterior, neste trabalho será desenvolvida uma abordagem para a RI que diferencia-se de outras abordagens por utilizar a opinião explícita dos usuários sobre a autoridade cognitiva de cada fonte de informação do SRI. Dessa forma, faz-se necessário comparar essa abordagem com outras abordagens já consagradas na literatura de RI na *Web*.

De acordo com (Buckley e Voorhees, 2004), as metodologias predominantes para avaliar SRIs são elaboradas utilizando o paradigma denominado de *Cranfield*[3] (Cleverdon, 1997). Nesse paradigma, um conjunto de dados de teste é utilizado para comparar a eficiência de diferentes abordagens e métodos para a RI. Os dados de teste consistem de um conjunto de sentenças sobre diversas necessidades de informação (o qual inclui as consultas para expressar tais necessidades de informação), um conjunto de documentos e um conjunto de julgamentos quanto à relevância dos documentos para cada uma das necessidades de informação. Manning et al. (2009) também destacam os elementos necessários para realizar a avaliação de um SRI, os quais incluem: i) uma coleção de documentos, ii) uma coleção de necessidades de informação, capazes de serem expressas como consultas e, iii) um conjunto de julgamentos a respeito da relevância de cada documento, com base nas necessidades de informação e nas consultas.

Quanto ao preparo das consultas a serem utilizadas na avaliação, pode-se observar o que é relatado por Webber (2010) e Bailey et al. (2003), quando dizem que há, com relação à essa tarefa, um conflito de requisitos difícil de satisfazer. Basicamente, há duas

[3]O nome *Cranfield* é utilizado em referência à cidade de Cranfield, no Reino Unido, onde originou-se a pesquisa sobre avaliação em RI que culminou no desenvolvimento de um conjunto de dados de testes pioneiro (Cleverdon, 1997), permitindo uma medida quantitativa precisa da eficiência de uma abordagem para RI (Harter, 1996). No entanto, o fato é que Cleverdon (1997) inicialmente batizou esse conjunto de dados de testes de forma diferente (Sanderson, 2010). Esse conjunto de dados de testes consiste de 1398 resumos de artigos sobre aerodinâmica, 225 consultas e uma coleção de julgamentos de relevância de todos os pares $(consulta, documento)$ (Manning et al., 2009).

formas de preparar o conjunto de tópicos (necessidades de informação e consulta) a serem utilizados pelos usuários participantes para a avaliação dos resultados de busca. Uma delas ocorre quando os tópicos são escolhidos aleatoriamente a partir de uma coleção de consultas reais (as quais podem ser obtidas, por exemplo, a partir de um *log* de consultas), enquanto a outra forma ocorre quando um usuário elabora uma consulta com base na sua necessidade de informação e então avalia os documentos a partir de tal consulta.

Essas duas formas de se preparar o conjunto de tópicos introduzem na avaliação vieses cujas naturezas são distintas. A abordagem em que os tópicos são escolhidos aleatoriamente a partir de um conjunto de consultas é interessante pois os tópicos são escolhidos com base em uma amostragem verdadeiramente aleatória desse conjunto. No entanto, quando as consultas são obtidas a partir de um *log* de consultas, por exemplo, é necessário que os avaliadores recriem a necessidade de informação a partir de uma consulta escolhida. Esse fato pode introduzir uma certa quantidade de artificialidade na avaliação realizada pelo participante sobre a relevância dos resultados de busca. Da mesma forma, na abordagem em que o avaliador exprime sua necessidade de informação por meio de uma consulta formulada por si mesmo, os tópicos não são escolhidos com base em uma amostragem aleatória, introduzindo uma margem de preferência nas consultas elaboradas, mas reduzindo a artificialidade das avaliações realizadas sobre a relevância dos resultados de busca (Webber, 2010).

A avaliação de SRIs visa principalmente estabelecer parâmetros para julgar qual abordagem para RI produz resultados de busca de maior *relevância*. Há uma vasta literatura que define o conceito de relevância e a sua relação com o conceito de qualidade de um item de informação (Barry, 1994; Goffman, 1964; Greisdorf, 2000; Kagolovsky e Mohr, 2001; Kargar, 2011; Knight e Burn, 2005; Lachica et al., 2008; Saracevic, 1975, 2007). Saracevic (1975), em um artigo seminal sobre o assunto no contexto da ciência da informação, discute a questão da relevância sob um arcabouço que considera a "comunicação de conhecimento" como um indicativo de relevância de uma informação[4]. Para esse autor, em seu sentido fundamental, a relevância é considerada como uma medida da efetividade do processo de comunicação que ocorre no contato com uma informação. O mesmo autor atualiza suas idéias em (Saracevic, 2007) e defende o ponto de vista de que "*ninguém precisa explicar o que é relevância a usuários de SRIs (. . .) as pessoas entendem a relevância intuitivamente.*" No entanto, ao definir formalmente a noção de relevância, o autor defende que esse conceito é formado por diversas dimensões, dentre as quais inclui a dimensão de "medida" (*measure*), sugerindo que a noção de relevância envolve uma avaliação gradual da efetividade do processo de comunicação citado.

[4]O termo *informação* está sendo utilizado em referência a um item de informação.

A relevância, no contexto de RI, é também conceituada como a medida com que a informação é transmitida por um documento dada uma consulta (Goffman, 1964). Isto é, a relevância é avaliada de acordo com uma necessidade de informação, de forma que um documento é relevante se contém informação que satisfaça determinada necessidade de informação. Além disso, a visão de relevância é geralmente desdobrada em duas (Kagolovsky e Mohr, 2001; Lachica et al., 2008): a relevância de tópico (*topical relevance*) e a relevância centrada no usuário (*user-centered relevance*). A primeira é objetiva e preocupada com a terminologia, podendo ser julgada por especialistas em determinado tópico. A segunda é subjetiva e depende da visão do usuário sobre a informação.

Para Lachica et al. (2008), a relevância de uma informação está relacionada à satisfação da necessidade de informação de um usuário ou o que esse usuário percebe dessa informação. Os autores exemplificam esses conceitos com uma história a respeito de um pescador procurando informação sobre material de pesca. Uma informação sobre a venda de um equipamento de pesca poderia ser *relevante* para o pescador, pois nesse caso é exatamente o que o pescador procura. Uma notificação sobre a chegada de um furacão seria considerada *importante*. Todas essas informações não seriam válidas se não satisfizerem critérios básicos de *qualidade*, tais como a veracidade e a completude. Os autores definem os conceitos de *qualidade* e *importância* da informação da seguinte forma[5]:

> A qualidade reflete o valor intrínseco de uma informação. Uma informação que não seja confiável ou que seja impossível de entender possui menos valor, mesmo sendo altamente relevante ou importante (...). A importância reflete o valor de uma informação do ponto de vista de um contexto mais amplo (...)

A questão da *Qualidade da Informação* (QI) relaciona-se com o conceito de relevância a medida que verifica diversas dimensões a respeito da informação, do ponto de vista de quem utiliza essa informação. Knight e Burn (2005) realizam uma extensa revisão sobre os autores que apresentam as dimensões e as classificações da QI. As autoras definem relevância como uma das dimensões da QI, a qual está relacionada à extensão na qual a informação é aplicável e útil para a tarefa sendo realizada. Essa definição tem claras relações com o que foi dito a respeito da relevância – um documento é relevante se atende à necessidade de informação (ou tarefa) de um usuário. Kargar (2011) analisou dados produzidos em *blogs* e considerou que a QI possui várias dimensões (ou critérios), as quais incluem a autoria, a disponibilidade, a completude, a representação, a compreensão e a objetividade, dentre outras. Wang e Strong (1996) realizaram um estudo sobre a qualidade

[5]Tradução livre do autor

de dados, ligados ao contexto de banco de dados, e concluiram que a qualidade também está relacionada com várias dimensões as quais captam a importância da informação para seus usuários, tais como o contexto, a representação e a acessibilidade. Barry (1994) identificou 23 categorias de critérios de relevância por meio de entrevistas à 18 estudantes universitários, os quais foram sub-divididos em sete grupos, dentre os quais estão os grupos de critérios pertencentes ao conteúdo do documento, à experiência prévia do usuário, à situação (contexto) do usuário, às preferências e crenças do usuário e à fonte dos documentos. Dentro de cada um desses grupos, Barry (1994) identificou alguns critérios utilizados pelos participantes para julgarem as informações com as quais interagiram (esses critérios foram também considerados neste trabalho), dentre os quais destacam-se: a extensão do conhecimento do usuário sobre a fonte e/ou o conteúdo da informação, a habilidade para compreensão do documento e a a extensão na qual o usuário concorda com a informação (o qual Barry (1994) chamou de "validade subjetiva").

No contexto da avaliação em RI duas métricas são bastante utilizadas para predizer a relevância e a qualidade das informações recuperadas por uma determinada abordagem para RI, as quais são chamadas de precisão (*precision*) e cobertura (*recall*). Essas duas métricas estão denotadas nas Equações 2.6 e 2.7 respectivamente. A precisão é a proporção de documentos recuperados que são relevantes, enquanto a cobertura é a proporção de documentos relevantes recuperados (Salton, 1971).

$$\mathrm{P} = \frac{\text{número de documentos relevantes recuperados}}{\text{número de documentos recuperados}} \tag{2.6}$$

$$\mathrm{C} = \frac{\text{número de documentos relevantes recuperados}}{\text{número de documentos relevantes}} \tag{2.7}$$

Utilizando essas métricas para avaliar diferentes abordagens para a RI é possível identificar a efetividade dessas abordagens por meio de uma análise quantitativa dos documentos relevantes recuperados. No entanto, essas duas métricas apresentam um *trade-off* natural entre elas, além de uma métrica ser mais importante do que outra dependendo da natureza da tarefa de RI (Manning et al., 2009). No entanto, uma discussão sobre esse assunto está além do escopo deste trabalho. Uma métrica para a avaliação de SRIs que relaciona e sintetiza as métricas de precisão e cobertura é chamada de *F-measure*, demonstrada na Equação 2.8, a qual representa uma média harmônica ponderada sobre as métricas de precisão e cobertura, sendo $0 \leq \alpha \leq 1$ e β parâmetros que controlam um peso para a precisão e a cobertura.

$$\mathrm{F} = \frac{(\beta^2 + 1)PC}{\beta^2 P + C} \text{ onde, } \beta^2 = \frac{1 - \alpha}{\alpha} \tag{2.8}$$

As métricas de precisão, cobertura e *F-measure* são adequadas quando se utiliza abordagens para a RI sobre um conjunto fechado de documentos, nos quais a ordem dos resultados não é importante, por exemplo quando se utiliza somente o modelo booleano para recuperar informação. No entanto, em muitas circunstâncias é necessário estender essas (ou definir novas) métricas de forma que elas possam lidar com resultados *ordenados* de busca. Nesse contexto uma métrica utilizada é a *Precision at K* – Precisão em K, a qual permite escolher um número fixo k de resultados de busca para serem avaliados sobre os quais se calcula a precisão somente em posições específicas no *ranking*. Um valor de k bastante utilizado é $k = 10$, situação na qual a métrica é denotada por P(10). Nesse caso, é contado o número de documentos relevantes nos 10 primeiros documentos na lista de documentos retornados em uma consulta (Manning et al., 2009). No entanto, Buckley e Voorhees (2004) e Sanderson (2010) comentam que essa métrica não é adequada para comparar diferentes métodos de RI, pois o único fato que discrimina os documentos retornados é a entrada ou a saída dentre os 10 primeiros resultados. Além disso, uma constante de corte de valor 10 representa uma variação muito grande nos resultados para diferentes tópicos. Esse fato é ilustrado pelo exemplo descrito em (Sanderson, 2010). Considerando dois tópicos distintos e um índice com $n = 100.000$ documentos, cuja quantidade de documentos relevantes para os referidos tópicos sejam respectivamente 100 e 10.000. A quantidade de documentos relevantes entre os 10 primeiros resultados para esses dois tópicos deverá ser bastante diferente, visto que a probabilidade de um documento relevante estar entre os 10 primeiros resultados é diferente para os dois tópicos. Outra métrica utilizada quando os resultados da abordagem para RI são apresentados em ordem é a *Mean Average Precision* (MAP), a qual compreende aspectos orientados à precisão e à cobertura. Seu cálculo é baseado na média da precisão para os k primeiros resultados de busca, os quais são calculados para vários tópicos.

Todas as métricas citadas como exemplo utilizam uma noção binária com relação à relevância dos documentos: um documento é relevante ou não-relevante, dada uma consulta. Alguns autores questionam essa escala binária de relevância e demonstram que os usuários de SRI costumam utilizar uma escala n-ária de relevância sobre os recursos (Borlund, 2003; Janes, 1991, 1994; Kagolovsky e Mohr, 2001; Schamber, 1994). Dessa forma, faz-se necessário também métricas que possam lidar com escalas n-árias de relevância dos recursos. Nesse sentido, algumas métricas vêm se popularizando principalmente para avaliar SRIs na *Web*, situação na qual o conjunto de documentos consultados é desconhecida. Estas métricas são baseadas no conceito de *Cumulated Gain* (CG) – Ganho Cumulativo (Chen e Zhang, 2009; Jarvelin e Kekalainen, 2002; Manning et al., 2009; Valizadegan et al., 2000; Yilmaz et al., 2008). O CG é a soma do ganho (isto é, do

valor aferido na escala de relevância utilizada para julgamento das informações) de cada documento do *ranking*, desde o documento apresentado na posição 1 até o documento apresentado na posição n. O CG de um documento que aparece na posição i de um *ranking* é calculado como a soma do CG dos documentos posicionados desde a posição 1 até a posição $i-1$. A Equação 2.9 ilustra esse cálculo, onde G[i] representa o *Gain* (ganho) do documento i. O ganho de um documento é geralmente calculado como o grau aferido à relevância de um documento de acordo com a escala utilizada (por exemplo, em uma escala binária o ganho de um documento poderia ser o valor 0 ou o valor 1).

$$\mathrm{CG}[i] = \begin{cases} \mathrm{G}[1], & \text{se } i = 1 \\ \mathrm{CG}[i-1] + \mathrm{G}[i], & \text{caso contrário} \end{cases} \tag{2.9}$$

Jarvelin e Kekalainen (2002) apontam que, utilizando uma métrica baseada no Ganho Cumulativo, os números em que são transformadas as avaliações de relevância dos documentos refletem as seguintes características com relação aos resultados de busca:

- Documentos altamente relevantes devem possuir valor de ganho maior do que documentos pouco relevantes;
- Quanto maior a posição no *ranking* para um documento relevante, menor é o valor para o usuário desse documento, pois a probabilidade do usuário avaliar o documento é menor, devido ao esforço, ao tempo e ao acúmulo de informação dos documentos anteriormente vistos.

A primeira característica é capturada por meio da utilização do próprio CG, no entanto a segunda característica é modelada utilizando a métrica denominada de *Discounted Cumulated Gain* (DCG), na qual uma função de desconto é utilizada para reduzir progressivamente o valor de CG no documento conforme sua posição i no *ranking* aumenta. A Equação 2.10 denota o cálculo da métrica de DCG, sendo b a base do $\log_b$, o qual representa a persistência do usuário para examinar os b primeiros resultados de busca. O cálculo dos valores de CG e DCG geram vetores denominados de $\vec{V}$, os quais possuem em cada dimensão i o valor de (D)CG do i-ésimo documento do *ranking*.

$$\mathrm{DCG}[i] = \begin{cases} \mathrm{CG}[i], & \text{se } i < b \\ \mathrm{DCG}[i-1] / \log_b i, & \text{se } i \geq b \end{cases} \tag{2.10}$$

Pode-se perceber que a métrica de DCG baseia o fator de desconto de seu valor no *ranking* de cada documento. Quanto maior o *ranking*, menor é a contribuição do ganho do documento para o ganho acumulado. Por exemplo, $\log_2 2 = 1$ e $\log_2 1024 = 10$. Isso

mostra que um elemento na posição 1024 do *rank* contribui com um décimo do ganho para o valor do ganho total acumulado, enquanto um elemento na posição 2 contribui com o valor integral de ganho (Manning et al., 2009). É claro também que não se pode aplicar o fator de desconto na primeira posição, pois $\log_2 1 = 0$ e nem em documentos que estejam em posições abaixo da base do $\log_b$, pois isso daria um *boost* no valor de DCG desses documentos.

Jarvelin e Kekalainen (2002) afirmam que é difícil avaliar a diferença de magnitude entre curvas de (D)CG geradas por duas abordagens para RI diferentes, fazendo com que não se possa realizar um teste de significância quando se utiliza essas duas métricas. Dessa forma, é necessária a definição de uma terceira métrica, a qual é baseada em uma curva ótima de (D)CG. Essa terceira métrica é chamada de *Normalized Discounted Cumulated Gain* (NDCG) e é calculada a partir dos vetores normalizados de (D)CG. Um vetor normalizado de (D)CG, é calculado por meio da normalização dos valores de (D)CG dos resultados de busca. Define-se, assim, um vetor chamado de "vetor ideal de ganho" denotado por V', o qual é definido como o vetor contendo os valores de ganho ordenados de forma decrescente. Por exemplo, um resultado de busca com 10 documentos poderia ter sido avaliado com os seguintes valores de ganho (G) $V = \{3, 2, 3, 0, 0, 1, 2, 2, 3, 0\}$, sendo que $V[i]$ representa o ganho do i-ésimo documento no *ranking* apresentado como resultado da busca (caso fosse necessário o cálculo da métrica de NDCG esses valores seriam os valores de DCG). Dessa forma, temos $V' = \{3, 3, 3, 2, 2, 2, 1, 0, 0, 0\}$.

Com base nessas definições, calcula-se o valor de N(D)CG para cada documento recuperado por meio da Equação 2.11. Nessa equação o tamanho dos vetores $\vec{V}$ e $\vec{V'}$ é igual a n.

$$norm\text{–}vec(\vec{V}, \vec{V'}) = \{v_1/v'_1, v_2/v'_2, ..., v_n/v'_n\} \tag{2.11}$$

A fim de avaliar um SRI utiliza-se os vetores *norm–vec* gerados por meio de cada abordagem de RI que se deseja avaliar. Por exemplo, considerando duas abordagens para RI denominadas de A1 e A2, geram-se vetores denominados respectivamente de $norm\text{–}vec_1$ e $norm\text{–}vec_2$ para cada abordagem. A partir desses vetores pode-se aplicar um teste de significância estatística (Baeza-Yates e Ribeiro-Neto, 1999; Manning et al., 2009; Smucker et al., 2007) a fim de verificar o grau de confiabilidade dos resultados.

É conhecido que na metodologia tradicional de avaliação de SRIs há vieses naturais relacionados à diferentes aspectos da avaliação. Por exemplo, Schamber (1994) destaca 80 fatores que influenciam o julgamento da relevância de documentos por parte de usuários avaliadores. Sendo assim, é necessária uma forma matemática de verificar se a diferença

entre os dados gerados pelas diferentes abordagens para RI são oriundos desse ruído ou se são de fato oriundos da opinião dos avaliadores das informações. Com base nessa questão são utilizados diferentes testes de significância estatística a fim de verificar os resultados de uma avaliação. Smucker et al. (2007) afirmam que o teste *T de Student* (teste t_{STAT}) é um teste apropriado para comparar métricas baseadas em média (como é o caso do NDCG). O valor de t_{STAT} para cada um dos vetores *norm–vec* gerados na avaliação é baseado na proporção entre a diferença das médias populacionais e amostrais e a incerteza padrão estimada (Currell e Dowman, 2009). A Equação 2.12 ilustra esse valor, na qual $\overline{x}$ é a média amostral, μ_0 é a média populacional, s é o desvio padrão e n é o tamanho da amostra.

$$t_{\text{STAT}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{2.12}$$

Para utilizar o valor de t_{STAT} deve-se gerar uma hipótese nula sobre os dados, a qual geralmente é a afirmação de que os dados gerados pelas duas abordagens geram os mesmos resultados em termos da métrica utilizada. A partir do cálculo do valor de t_{STAT}, compara-se esse valor com o valor de t_{CRIT} (o qual é obtido por meio de uma tabela) e, caso $|t_{\text{STAT}}| \geq t_{\text{CRIT}}$, a hipótese nula pode ser rejeitada. Dessa forma, o valor de t_{CRIT} representa o limite mínimo para o qual o valor de t_{STAT} deve ser calculado como resultado do teste de significância. Ainda, é possível aplicar dois tipos de teste de significância, a saber: o teste bicaudal, no qual o efeito da hipótese nula opera em ambas os lados da curva normal; e o teste monocaudal, no qual o efeito da hipótese nula opera em um dos lados da curva. Por exemplo, quando a hipótese é de que a média de duas amostras sejam *diferentes*, usa-se o teste bicaudal, enquanto quando a hipótese é de que a média de uma amostra seja *maior* (ou menor) do que a média da outra amostra, usa-se o teste monocaudal.

Neste trabalho foi utilizada a métrica de NDCG a fim de comparar a abordagem para RI desenvolvida com uma abordagem tradicional para a RI de recursos na *Web*. Além disso, a fim de realizar um teste de significância sobre os dados gerados na avaliação foi utilizado o teste T_{STAT} monocaudal, uma vez que a hipótese nula operava sobre a melhoria, ou não, entre as abordagens para RI (isto é, a hipótese nula afirmava que o valor de NDCG de uma abordagem é maior do que o valor de outra). No próximo capítulo será apresentada uma discussão sobre o conceito de *Folkauthority*, o qual serviu de arcabouço para o desenvolvimento da abordagem para RI proposta neste trabalho.

Capítulo

3

Folkauthority

Fokauthority (folk + authority) é um neologismo proposto por Pereira e da Silva (2008b) a fim de designar a concessão de autoridade cognitiva por meio de Folksonomia. Esse conceito está relacionado com a teoria de Wilson (1983) sobre Autoridade Cognitiva, uma teoria da epistemologia social que explica o tipo de autoridade que influencia os pensamentos que as pessoas conscientemente julgam serem adequados para serem utilizados. Wilson (1983) diferencia a autoridade cognitiva da autoridade administrativa – o tipo de autoridade que emerge de uma posição hierarquia imposta. A autoridade cognitiva representa a influência que uma autoridade pode causar no pensamento de outro indivíduo. Esse tipo de autoridade define "quem sabe o quê sobre o quê". O autor em sua obra descreve o processo de aquisição de conhecimento de "segunda mão" por meio de autoridades cognitivas. O conhecimento de segunda mão é aquele adquirido por meio da consulta à outras pessoas, diferente daquele conhecimento adquirido por meio da experiência ou da verificação própria. A discussão do autor refere-se também à questões relacionadas ao controle da qualidade das informações passíveis de serem adquiridas/recuperadas.

Wilson (1983) discute que a autoridade cognitiva pode ser atribuída não somente à indivíduos, mas também à livros, instituições ou à qualquer entidade que possa servir de fonte de informação. Além disso, o autor relaciona algumas questões relevantes sobre a autoridade cognitiva, as quais são consideradas na abordagem de *Folkauthority*, explicada com maiores detalhes adiante. Esse tipo de autoridade sempre envolve pelo menos duas entidades: a autoridade (indivíduo, livro, instituição, etc.) e o indivíduo que reconhece

essa autoridade. Por exemplo, uma pessoa pode possuir grande conhecimento em determinado assunto, no entanto ela pode não ser reconhecida por outras pessoas como uma autoridade cognitiva. Dessa forma, a autoridade cognitiva depende do *reconhecimento* (ou da concessão, como será referenciada no contexto de *Folkauthority*) por parte de alguém. Essa afirmação leva à outra suposição feita por Wilson (1983) sobre a autoridade cognitiva – a de que esse tipo de autoridade está sempre relacionada à algum âmbito de interesse, de forma que uma entidade possa ser considerada autoridade em determinados assuntos, enquanto em outros não exista esse mesmo reconhecimento. Além disso, a autoridade cognitiva possui níveis de reconhecimento, de forma que uma entidade possa ser reconhecida com muita ou pouca autoridade cognitiva.

Ao analisar parte da literatura relacionada à Ciência da Informação, à Qualidade da Informação e à *Web*, é possível perceber que alguns autores já discutiram a relação entre autoridade (cognitiva), credibilidade e qualidade/relevância da informação disponível na *Web*. Kleinberg (1999) e Amento et al. (2000) propõem estimar a noção de "autoridade" de um recurso na *Web* por meio da análise das estruturas de *hiperlink*. Essa noção de autoridade considera que se um documento D_1 referencia outro documento D_2, então o autor do documento D_1 considera D_2 um documento relevante ou de qualidade. Amento et al. (2000) propõem uma análise da estrutura dos *links* e reportam bons resultados em termos de Precisão em K em um teste com dados de um diretório *Web*. Fritch e Cromwell (2001) desenvolveram um modelo para descrever como ocorre a concessão de autoridade cognitiva aos recursos na *Web* por parte dos usuários.

Rieh (2002) descreve que a principal preocupação com a questão da autoridade e da qualidade da informação, no contexto da Ciência da Computação, é com relação à efetividade dos mecanismos de busca. A autora realiza um estudo sobre o julgamento da autoridade cognitiva e da qualidade de recursos na *Web* e relaciona os principais fatores que influenciam este julgamento. Wathen e Burkell (2002) avaliam como a credibilidade da informação é aferida e percebida na *Web* e afirmam que a credibilidade da informação influencia as atitudes, o comportamento e o conhecimento do usuário.

McKenzie (2003) explica a questão da autoridade cognitiva sob o ponto de vista do construcionismo, o qual considera que o conhecimento é construído por meio do diálogo e da ação concreta e não somente por meio da mente das pessoas. Segundo a autora "o foco do construcionismo é na linguística e não nos processos cognitivos".

Russel (2005) propõe um sistema que permite a descoberta e a definição de autoridades cognitivas – nesse caso a autoridade é relacionada aos próprios usuários do sistema e não aos recursos disponíveis. A fim de colaborar com esses requisitos, é proposto um sistema que permite o uso de *tags* para descrever a autoridade cognitiva dos usuários, o qual Russel

(2005) denomina de *Contextual Authority Tagging*. Diferente da abordagem de *Folkauthority*, nesse sistema as entidades, as quais são reconhecidas como autoridades cognitivas, não são acompanhadas de nenhuma informação de autoria/disponibilização própria, isto é, enquanto a abordagem de *Folkauthority* trata a questão da concessão de autoridade cognitiva como uma metacategorização[1], a abordagem de *Contextual Authority Tagging* não menciona a consideração de autoridades como fonte de informação. Dessa forma, a abordagem de *Folkauthority* é relacionada com a questão de recuperar informação de qualidade/relevância, enquanto os requisitos relacionados ao sistema *Contextual Authority Tagging* não atingem tais méritos.

Savolainen (2007) estima a percepção da confiança e da credibilidade da informação e como essa percepção está relacionada a seleção de fontes de informação. Metzger (2007) apresenta um modelo descrevendo as habilidades dos usuários para avaliar a credibilidade de recursos na *Web*.

Baseado nesses trabalhos, é desenvolvida a idéia de permitir que os usuários possam explicitamente informar a autoridade cognitiva das fontes de informação de um SBF por meio de *tags*, além de utilizar o resultado dessa atividade para melhorar a RI. Essa abordagem foi chamada de *Folkauthority* (Pereira e da Silva, 2008b), a qual será explicada com maiores detalhes nas seções seguintes.

3.1 Cadeia de autoridades

A definição da abordagem de *Folkauthority* diz respeito à concessão de autoridades cognitivas utilizando a estrutura provida pelo conceito de Folksonomia. O trabalho de Russel (2005) demonstrou que a técnica de *tagging* é uma ferramenta útil para denotar a autoridade cognitiva de uma fonte de informação. Esse autor apresentou um SBF no qual é possível atribuir *tags* aos próprios usuários a fim de descrever suas competências e suas habilidades cognitivas, na opinião de quem atribui tais *tags*. Pereira e da Silva (2008b) consideram que nos SBFs as fontes de informação são os próprios usuários do sistema, uma vez que esses usuários são responsáveis por publicar/disponibilizar e categorizar as informações. Os autores sintetizam as afirmações de Russel (2005) e Wilson (1983) para propor a abordagem de *Folkauthority*, a qual é baseada na expectativa de que utilizar a técnica de Folksonomia para descrever a autoridade cognitiva das fontes de informação pode auxiliar na tarefa de RI em sistemas que se baseiem nessa abordagem.

[1]Cada usuário ao qual é concedida autoridade é acompanhado de suas categorizações, pois são vistos como fontes de informação.

Em sua obra, Wilson (1983) afirma que uma autoridade cognitiva em determinado assunto possui maior probabilidade de desenvolver esquemas de categorização adequados para as informações, de possuir itens de informação de melhor qualidade e de manter contatos com outras autoridades nesse assunto. Esse mesmo autor defende que a autoridade cognitiva é responsável também pelo controle da qualidade de informações recuperadas/adquiridas. Uma vez que em um SBF os documentos são indexados a partir das *tags* – isto é, dos esquemas de categorização elaborados pelos próprios usuários – a efetividade do resultado da tarefa de RI nesses sistemas depende também da "qualidade" com que os esquemas de categorização são elaborados. Sendo assim, Pereira e da Silva (2008b) argumentam que a abordagem de *Folkauthority* pode auxiliar no processo de RI em SBFs.

Para exemplificar a utilização da abordagem de *Folkauthority* é possível citar o sistema proposto por Pereira e da Silva (2008b), o qual é denominado de *Cognitive Authority on the Web (CAW)*. Esse sistema integra recursos de outros SBFs (tais como o *Delicious* e o *Flickr*) de forma que os recursos do sistema *CAW* sejam os próprios recursos dos SBFs integrados. No sistema *CAW* é possível atribuir *tags* aos próprios usuários, no entanto, diferente das *tags* aplicadas aos recursos, as quais são utilizadas para descrever sobre o que esses recursos tratam, as *tags* aplicadas aos usuários descrevem a autoridade cognitivas desses. Sendo assim, nesse sistema há uma meta-categorização: os objetos são categorizados pelos usuários que, por sua vez, são também categorizados de acordo com sua autoridade cognitiva. Há uma categorização das informações em um primeiro nível, a qual é realizada pelos usuários do sistema, e uma categorização dos usuários (fontes/publicadores de informação) em um segundo nível, a qual também é realizada pelos próprios usuários do sistema

Em um SBF que adote a abordagem de *Folkauthority*, é gerada uma rede social formada pelos usuários concessores de autoridade cognitiva[2] e pelas próprias autoridades cognitivas, sendo que cada par de nós nessa rede social refere-se à uma concessão de autoridade cognitiva. Dessa forma, pode-se pensar nessa rede contendo arestas rotuladas com informações sobre a *tag* e um peso associado a autoridade, as quais descrevem um nível de reconhecimento da autoridade cognitiva da fonte de informação (Wilson, 1983).

Uma *rede de autoridades* é composta por todas as concessões de autoridade realizadas por um usuário. Por exemplo, suponha um usuário denominado de A o qual tenha categorizado o usuário denominado de B com a respectiva *tag* e peso: $\{information,3\}$, a

[2]O termo usuário concessor refere-se àquele usuário que tenha concedido autoridade à outro usuário. Conceder autoridade, no contexto de *Folkauthority* (Pereira e da Silva, 2008b) significa aplicar uma *tag* à uma fonte de informação a fim de descrever sua autoridade cognitiva.

qual descreve a autoridade desse usuário. A partir da ocorrência desse fato, o usuário B passa a fazer parte da rede de autoridades do usuário A, a qual é composta pelo conjunto de autoridades categorizadas por esse usuário. Sendo assim, essa relação pode ser modelada como um grafo dirigido (unidirecional). A utilização da abordagem de *Folkauthority* origina outra rede social, chamada de *cadeia de autoridades*, a qual é definida a partir da justaposição das diversas redes de autoridades (Pereira e da Silva, 2008b). Desta cadeia de autoridades podem ser extraídas diversas informações úteis para a integração à um esquema de *scoring* para RI. Por exemplo, Russel (2005) e Pereira e da Silva (2008b) discutem algumas informações que podem ser extraídas da cadeia de autoridades, as quais incluem: i) a descoberta de autoridades, isto é, a possibilidade de descobrir quem é uma autoridade em determinado assunto, ii) o reconhecimento de competências, o qual permite identificar as competências cognitivas de determinada entidade (Pereira e da Silva, 2008b), iii) a utilização de uma medida de "peso" para cada autoridade, de acordo com sua popularidade. Dentre as questões discutidas pelos referidos autores é possível destacar a afirmação de que, no momento de se recuperar informação, pode-se priorizar informação que foi disponibilizada por uma autoridade no assunto contido na consulta.

A Figura 3.1 ilustra uma parte da cadeia de autoridades, na qual é possível observar a concessão de autoridade do usuário A para o usuário B, que por sua vez concede autoridade ao usuário C com a *tag*/peso {`information`,3}. O usuário C, por sua vez, concede autoridade à mais dois usuários nas *tags*/pesos {`indexing`,2} e {`web`,2}. É possível observar na figura um destaque para as concessões de autoridade entre os usuários A, B, C e D, as quais foram realizadas utilizando-se a *tag* `information`. Esse destaque é uma referência à expressão "cadeia restringida pela *tag*", a qual será utilizada com frequência adiante neste trabalho. A noção de restringir a cadeia de autoridades por determinada *tag* diz respeito à considerar somente as concessões de autoridades realizadas utilizando-se essa *tag*, sendo que na Figura 3.1 a cadeia em destaque coincide com a cadeia restringida pela *tag* `information`.

Dentre as informações que poderiam ser extraídas da cadeia de autoridades, neste trabalho foi utilizado o cálculo do *PageRank* à Priori (White e Smyth, 2003) a fim de estimar o peso que cada autoridade exerce no *score* de um documento recuperado. O esquema de *ranking* que utiliza essa informação será descrito com maiores detalhes no Capítulo 4. No entanto, é importante esclarecer alguns princípios da abordagem para RI proposta neste trabalho.

Ao aplicar um algoritmo de *PageRank* à Priori na cadeia de autoridades, é possível estabelecer (considerando os pesos associados aos nós da rede) um número que sintetize uma noção de autoridade global na rede (Brin e Page, 1998) a qual considera, além da

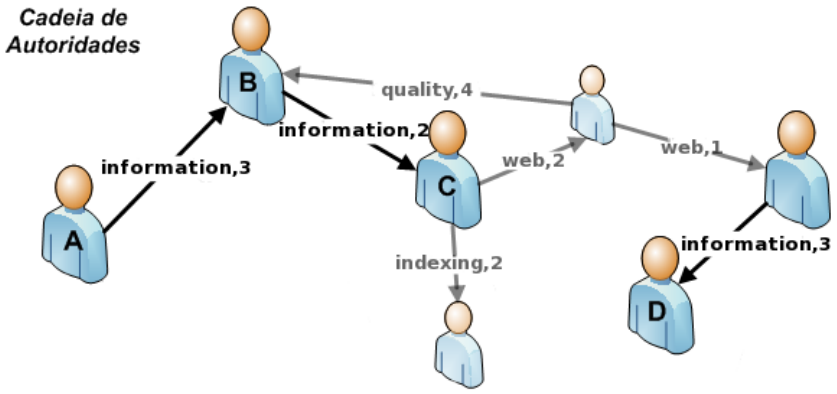

Figura 3.1: Cadeia de autoridades. Destaque para a cadeia restringida pela *tag* `information`

topologia da rede, os pesos concedidos às autoridades. Essa medida permite responder à pergunta "quais entidades são mais importantes na rede com relação a um outro indivíduo ou a um conjunto de indivíduos ?" (White e Smyth, 2003). Partindo da premissa de que as informações categorizadas/disponibilizadas pelas autoridades são mais "importantes", então é natural pensar em utilizar tal medida a fim de diferenciar o *ranking* dos documentos categorizados por autoridades.

Para clarear essa idéia, é possível exemplificar um sistema que utilize a abordagem de *Folkauthority* e que diferencie o *score* dos documentos recuperados a partir do valor de *PageRank* à Priori das autoridades. Considere uma consulta contendo os termos `information retrieval interaction`. O primeiro passo a ser realizado será restringir a cadeia de autoridades de forma que a sua topologia reflita a concessão de autoridades contendo *tags* com os termos `information`, `retrieval` e `interaction`, de acordo com os operadores booleanos expressos na consulta. Deste modo, será possível calcular o valor do *PageRank* à Priori de cada autoridade contida na rede restrita. Uma vez calculado esse valor para cada autoridade, pode-se utilizar um esquema de *ranking* no qual o cálculo do *score* de cada documento leve em consideração o valor de *PageRank* à Priori do usuário que o categorizou. Caso o documento tenha sido categorizado por um usuário que não é autoridade em nenhum dos termos da consulta, o esquema de *ranking* deve poder lidar com essa situação, de forma que ainda assim haja uma definição de importância do documento para a consulta.

A abordagem que será discutida é baseada em um incremento no *score* de cada documento. Esse incremento será calculado a partir do valor de *PageRank* à Priori de cada autoridade que categorizou um recurso contendo termos da consulta, sendo que somente as

autoridades categorizadas com *tags* relacionadas à consulta contribuem para o incremento dos documentos. Além disso, o incremento incide sobre o valor de um esquema de cálculo de *score* tradicional, tal como o *tf-idf* (Singhal, 2001). A abordagem de *Folkauthority* para categorizar cada fonte de informação de acordo com sua autoridade cognitiva pode também ser utilizada para encontrar autoridades em um assunto e para conhecer as competências de um usuário. Uma vez que as autoridades são categorizadas utilizando o mesmo esquema que os documentos (isto é, por meio de *tags*), as discussões realizadas sobre a RI e a abordagem de *Folkauthority*, as quais serão tecidas adiante, poderão naturalmente serem estendidas à recuperação de autoridades em determinado assunto (ao invés de um documento em determinado assunto, conforme discutido neste trabalho). Apesar da abordagem apresentada poder ser naturalmente estendida para a recuperação de autoridades, não é possível afirmar, sem que haja uma melhor investigação, que a recuperação das *informações publicadas/disponibilizadas pelas autoridades* gere resultados semelhantes à recuperação de *autoridades em determinado assunto*. Uma vez que as *tags* utilizadas para categorizar as autoridades são diferentes das *tags* utilizadas para categorizar os documentos, os *resultados* com a utilização da abordagem apresentada neste trabalho não podem ser estendidos para a recuperação de autoridades em determinado assunto, no entanto a abordagem em si pode ser estendida quase que de forma trivial.

Na próxima seção será discutida com maiores detalhes as lacunas existentes entre o conceito de *Folkauthority* e a RI, as quais geram questões de pesquisa relacionadas à este trabalho.

3.2 Lacunas entre o Folkauthority e a RI

Um trabalho que de fato avaliou a relação entre o conceito de *Folkauthority* e a RI foi (Pereira e da Silva, 2008c), no qual foram discutidos vários benefícios com relação ao conceito de *Folkauthority* e a sobrecarga de informação. O autor realiza um estudo a fim de validar a hipótese de que "*... a aplicação do conceito de autoridade cognitiva por meio de folksonomia eleva a precisão da informação recuperada e ameniza o impacto da sobrecarga de informação nestes sistemas.*" Esse estudo é baseado em uma simulação da categorização dos documentos por parte dos usuários e em uma simulação da cadeia de autoridades. No entanto, é possível apontar algumas questões a serem consideradas na simulação realizada em (Pereira e da Silva, 2008c), a qual é baseada em cinco passos que definem o conjunto de elementos a serem utilizados no estudo, a saber: as *tags*, os usuários, os documentos, a categorização dos documentos e a concessão de autoridades cognitivas.

Na etapa de definição de *tags* foram definidas 250 *tags* as quais foram utilizadas em todo o estudo realizado por Pereira e da Silva (2008c). Na etapa de definição dos usuários, foram gerados 100 usuários e, para cada usuário foram associadas 20 *tags* a fim de denotar seu "vocabulário", isto é, de denotar as *tags* que foram utilizadas por cada usuário para categorizar os documentos. No entanto, as 20 *tags* associadas a cada um dos 100 usuários foram escolhidas aleatoriamente, o que diverge daquilo que é apontado na literatura como um modelo para a utilização de *tags* entre os usuários de SBFs (Golder e Huberman, 2006; Halpin et al., 2007; Trant, 2009). É conhecido que os termos utilizados tanto para descrever um documento quanto aqueles pertencentes ao vocabulário de um usuário possuem uma relação entre si. Um documento possui um *tema* ou *assunto* e dessa forma os termos mais importantes contidos no documento possuem uma relação com o significado desse assunto (Manning et al., 2009). Além disso, cada usuário possui um ou mais *assuntos de interesse* e os termos utilizados pelo usuários para descrevê-los também guardam essa mesma relação. Ainda assim, considera-se que a questão da distribuição da frequência de uso das *tags* de cada usuário foi bem procedida em (Pereira e da Silva, 2008c), pois nesse trabalho utilizou-se a distribuição de *Paretto* para simular a frequência de uso das *tags* por cada usuário.

No momento de definição das categorizações dos documentos, a simulação demonstrada em (Pereira e da Silva, 2008c) procede escolhendo um número fixo de 1000 documentos. Além disso, para cada documento foram escolhidos 20 usuários aleatoriamente, a fim de denotarem os usuários categorizadores dos documentos. A atribuição de *tags* aos documentos não obedece à uma distribuição de probabilidade na qual as *tags* relacionadas ao assunto do documento possuem maior probabilidade de serem utilizadas. Essa distribuição é "aleatória". No entanto, nessa simulação foram utilizadas entre 2 e 5 *tags* para descrever cada documento (para cada usuário). Esse é o número de *tags* de fato utilizadas pelos usuários para categorizar os documentos, conforme apontado na literatura que já investigou essa questão (Golder e Huberman, 2006). Por fim, a topologia da cadeia de autoridades foi definida de forma aleatória (cada usuário categorizava um número aleatório entre 0 e 10 de autoridades). No entanto, conforme apontado por autores na área de Análise de Redes Sociais, um modelo generativo aleatório não é o mais adequado para descrever uma rede social com características da cadeia de autoridades. Mika (2007), ao discorrer sobre o modelo generativo conhecido como modelo *random graph* afirma que uma rede gerada aleatoriamente não é a ideal para denotar redes com características da cadeia de autoridades. Em seu texto o autor diz que devido à limitações de espaço, é improvável que as relações sociais aconteçam de forma totalmente aleatória, apesar de algumas relações serem geradas ao acaso ou acidentalmente. Existe uma maior

probabilidade de que essas relações aconteçam em ambientes sociais limitados, sendo as relações ao acaso mais esporádicas. Além disso, Kleinberg (1999) aponta para um modelo de *hubs* (nós com muitas arestas de saída) e *authorities* (nós com muitas arestas incidentes) a fim de descrever uma rede de "autoridades" (influência) de páginas *Web*.

A Figura 3.2 mostra um excerto da cadeia de autoridades avaliada neste trabalho, a qual foi gerada por meio da mineração de uma rede social existente na *Web* (maiores detalhes sobre esse procedimento podem ser encontrados no Capítulo 4, Seção 4.2). Nessa figura é possível visualizar dois nós destacados pelos termos "*Authority*" e "*Hub*". Um nó *Authority* representa um usuário que publica/disponibiliza muita informação e é bastante reconhecido como autoridade em alguns assuntos. Já o nó *Hub* representa um usuário que realiza intensa atividade de categorização de autoridades cognitivas, podendo ou não ser também reconhecido como autoridade em alguns assuntos.

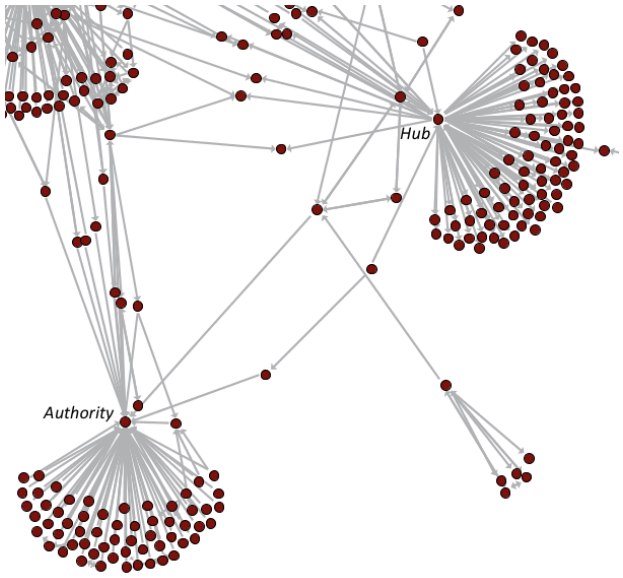

Figura 3.2: Topologia da cadeia de autoridades.

Ainda com relação ao trabalho de Pereira e da Silva (2008c), a abordagem apresentada para RI utilizando *Folkauthority* foi comparada com uma abordagem para RI baseada na ordem cronológica dos resultados. Isto é, todos os documentos categorizados com *tags* pertencentes à consulta foram ordenados de acordo com a data em que foram disponibilizados no sistema e então foram comparados com o resultado de busca gerado por uma abordagem com *Folkauthority*. É claro que esta não é uma abordagem tradicionalmente utilizada na *Web* (apesar de ser utilizada pelo sistema *Delicious* na época em que o estudo de Pereira e da Silva (2008c) foi publicado). Dessa forma, há uma lacuna sobre estudos na

área de RI e de *Folkauthority*, os quais demonstrem comparações entre uma abordagem para RI que utilize o conceito de *Folkauthority* e uma abordagem tradicional de RI na *Web* (como é o caso do esquema de *ranking tf-idf*). Além disso, carece também de investigações o problema de gerar dados que sejam baseados em modelos compatíveis com uma cadeia de autoridades (um grafo orientado cuja distribuição das arestas incidentes nos nós obedeça uma ordem de probabilidade conhecida). Uma possibilidade para o tratamento dessa questão é a geração da cadeia de autoridades a partir de dados de uma rede social real – abordagem a qual foi utilizada neste trabalho.

Neste trabalho, a utilização de uma rede social real para tratar a questão da definição da cadeia de autoridades permitiu obter a topologia da cadeia (quais usuários concederam autoridades e à quem), no entanto os dados relacionados à concessão de autoridade (quais *tags* e níveis foram atribuídos às autoridades) foram simulados com base no conjunto de *tags* dos dois usuários que participaram da concessão (as quais foram obtidas a partir de um sistema real). Além disso, utilizou-se o algoritmo de *PageRank* à Priori para calcular a importância de uma autoridade em uma *tag* dentro da cadeia de autoridades. Outra questão tratada neste trabalho foi a lacuna relacionada aos modelos para descrever a categorização dos documentos por parte dos usuários. Essa questão também foi tratada com base na captura de dados de um SBF real, conforme será descrito com detalhes no Capítulo 4.

3.3 Ferramentas Relacionadas

Há, na área da Ciência da Informação e da Ciência da Computação, vários autores que destacam a relação entre qualidade/relevância da informação e o conceito de autoridade (cognitiva) (Alexander e Tate, 1999; Barry, 1994; Janes, 1991; Katerattanakul e Siau, 1999; Knight e Burn, 2005; Rieh, 2002; Schamber et al., 1990). A autoridade está relacionada a influência de uma entidade em um contexto social, enquanto a autoridade cognitiva representa a influência dessa entidade com relação ao conhecimento. A proposta apresentada neste trabalho discute a relação entre a autoridade cognitiva e a RI. Dessa forma, é interessante apresentar alguns sistemas que exploram na prática a noção de autoridade para lidar com a informação produzida/consumida pelos próprios usuários. A maioria dos sistemas utiliza uma noção de autoridade calculada automaticamente a partir da indicação da relação social entre dois indivíduos em uma rede.

Apesar de tantos autores apontarem para a relação entre autoridade e qualidade, há poucos sistemas disponíveis publicamente que se aproveitam dessa relação de forma explícita a fim de buscar informação (com exceção dos populares mecanismos de busca

que exploram a estrutura de *hiperlink* e utilizam algoritmos baseados em autoridade para calcular a importância de uma página *Web*.). Há uma tendência atual em se utilizar os sistemas de redes sociais na *Web* para medir a influência dos usuários e em identificar os assuntos de interesse dos usuários a partir dessas redes sociais. Há também o interesse em redes sociais de co-autoria e um crescente interesse no conceito de redes sociais acadêmicas. Essa tendência é interessante pois uma rede acadêmica possui uma forte relação com o conhecimento dos seus usuários, como é o caso de uma cadeia de autoridades.

No sistema denominado de *Klout*[3], um usuário pode se cadastrar utilizando sua conta de um outro sistema, a saber o *Facebook*[4] ou o *Twitter*[5]. O sistema *Klout* prediz um valor para a influência de um usuário em um tópico a partir de informações geradas em diversos sistemas de redes sociais. No *website* do sistema *Klout* é possível obter a seguinte descrição sobre o cálculo do valor de influência dos usuários:

> *"The Klout Score measures influence based on your ability to drive action. Every time you create content or engage you influence others. The Klout Score uses data from social networks in order to measure: True Reach – How many people you influence, Amplification – How much you influence them, Network Impact – The influence of your network".* [6]

Como é possível observar, o sistema *Klout* prediz a influência que um usuário possui no sistema a partir das relações sociais que esse usuário possui. A Figura 3.3 mostra a interface do sistema na qual é possível encontrar usuários que sejam influentes em um assunto, no caso o assunto "cooking". Essa interface mostra o valor calculado para a influência de um usuário neste assunto e ordena os resultado com base nessa valor, sendo possível encontrar uma pessoa influente em um tópico e "segui-la" no *Twitter* ou "adicioná-la" ao *Facebook*. Outra funcionalidade que pode ser destacada no sistema *Klout* é a possibilidade de indicar explicitamente o reconhecimento da influência de outro usuário (na interface mostrada na Figura 3.3, essa funcionalidade pode ser utilizada com um clique na figura identificada com "+K". Apesar dessas funcionalidades estarem bastante relacionadas com a intenção de demonstrar uma concessão de autoridade entre os usuários, o sistema *Klout* não a faz por meio de *tags*, bem como não apresenta as informações disponibilizadas/publicadas pelos usuários na consulta (como faz o sistema estudado

[3]http://klout.com/

[4]http://www.facebook.com

[5]http://twitter.com/

[6]Tradução do autor: O *Klout Score* mede a influência baseada na sua habilidade para realizar ação. Toda vez que você cria um conteúdo você influencia outras pessoas. O *Score Klout* usa dados da redes sociais a fim de medir: O Alcançe – Quantas pessoas você influencia, A Amplificação – Quanto você influencia as pessoas, O Impacto da Rede – A influência da sua rede.

neste trabalho, o qual é apresentado no Capítulo 4, Seção 4.3), mas permite visualizar a atividade dos usuários nos sistemas integrados ao sistema *Klout* (por meio do link "*Best Content*").

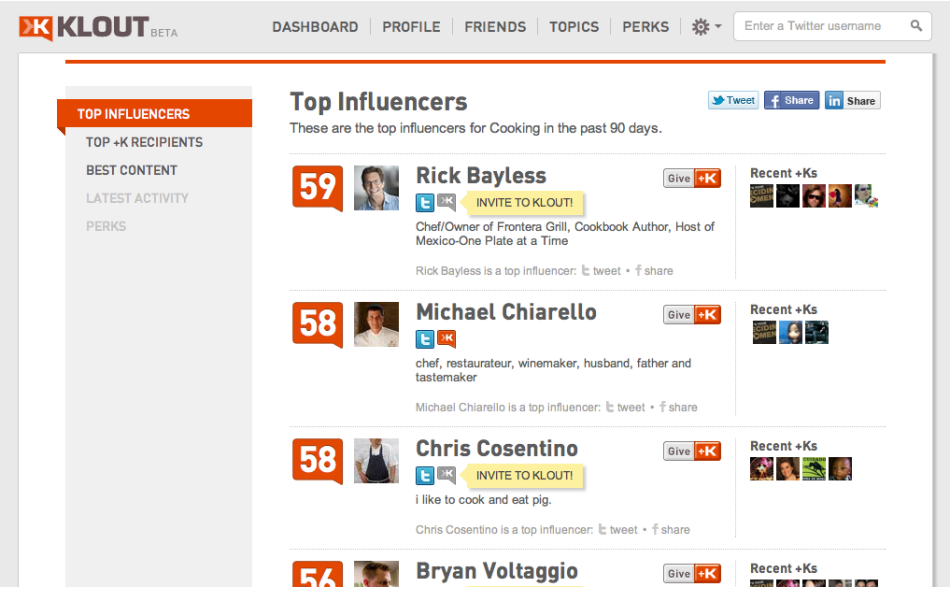

Figura 3.3: Interface para busca de usuários influentes em um tópico no sistema *Klout*.

Outros sistemas também utilizam essa mesma ideia e realizam a medição da influência e da autoridade de um usuário a partir de suas relações em sistemas de redes sociais. Por exemplo, o sistema *PeerIndex*[7] permite medir aquilo que, nos termos conceituais do sistema, é chamado de "capital social" e verificar os assuntos que cada usuário tem se interessado. O sistema *PostRank*[8] mede a atividade dos usuários e gera um valor que, nos termos do sistema, serve como "*indicador da relevância e da influência de um website, de uma história ou de um autor.*"[9]. O que pode ser notado nesse sistema é que o mesmo permite a busca de diferentes recursos. O sistema *Jive*[10] é utilizado como uma ferramenta para o monitoramento das atividades de usuários em determinado assunto nos sistemas de redes sociais. Todas essas aplicações possuem a característica de agregar vários sistemas e realizar uma mineração nas atividades sociais dos usuários a fim de poder recomendar ou recuperar algum tipo de informação (ou mesmo uma pessoa que possa ter uma informação), as quais são em geral voltadas (mas não restritas) às atividades da indústria, tais como a definição de "tendências" para o mercado.

[7] http://www.peerindex.com/
[8] http://www.postrank.com/
[9] Tradução livre do autor
[10] http://www.jivesoftware.com/

O sistema *LinkedIn*[11] – uma plataforma para rede social entre profissionais – recentemente lançou uma ferramenta capaz de encontrar pessoas com determinada habilidade ou especialidade (*skill* e *expertise*). Nessa ferramenta, os especialistas são indexados de acordo com aquilo que descreve seus perfis. Além disso, um profissional pode ser "recomendado" por outro profissional de forma deliberada, no sentido de ter as habilidade declaradas em seu perfil confirmadas. Apesar dessa não ser uma concessão de autoridade cognitiva, já que o usuário auto-declara suas habilidades, o sistema baseia sua busca nas habilidades relacionadas ao conhecimentos dos usuários, fato que se relaciona com a abordagem de *Folkauthority*. O sistema oferece então uma interface para consulta textual permitindo que um usuário descreva o assunto para o qual deseja encontrar especialistas.

A principal diferença entre os sistemas apresentados e a abordagem de *Folkauthority*, a qual foi apresentada neste capítulo, é que nessa última a autoridade é indicada explicitamente por meio de *tags*. A partir dessa atividade, realizada pelos próprios usuários do sistema, pode-se aferir um *ranking* das informações publicadas pelas autoridades (ou das próprias autoridades) a fim de realizar a tarefa de RI. Com base nessa afirmação, deseja-se avaliar se o *ranking* gerado a partir da indicação deliberada das autoridades cognitivas pode apresentar informação de melhor relevância e de melhor qualidade do que os *rankings* gerados por uma abordagem tradicional de RI. No próximo capítulo será apresentada a metodologia utilizada neste trabalho para avaliar a aplicação da abordagem de *Folkauthority* para RI, bem como será apresentada a modelagem de um sistema que utiliza o conceito de *Folkauthority* para recuperar informação disponibilizada pelas autoridades.

[11] http://www.linkedin.com/

Capítulo

4

Modelagem e Simulação

Baseado nas discussões anteriores este trabalho visa investigar como um SRI que faça uso do conceito de *Folkauthority* e de informações extraídas da cadeia de autoridades pode melhorar a qualidade e a relevância dos documentos recuperados. Nesta pesquisa considera-se a hipótese de que utilizar a opinião dos usuários sobre a autoridade cognitiva das fontes de informação em um esquema de *ranking* pode ajudar a identificar documentos que melhor satisfaçam determinada necessidade de informação, no sentido de serem apresentados documentos de maior relevância e qualidade nas primeiras posições. Essa hipótese é fundamentada na relação entre os conceitos de autoridade, qualidade e relevância (Pereira e da Silva, 2008b; Rieh, 2002).

Apesar de verificada a viabilidade da aplicação do conceito de *Folkauthority* para melhoria dos resultados de buscas em SBFs, conforme demonstrado em (Pereira e da Silva, 2008a,b), ainda restam questões acerca de como as característica intrínsecas em torno desse conceito podem ser exploradas no contexto de RI. Essas questões incluem, por exemplo, um modelo de RI a ser utilizado (Baeza-Yates e Ribeiro-Neto, 1999), os dados a serem calculados sobre a cadeia de autoridades em tempo de indexação e em tempo de busca (Zobel e Moffat, 2006), os ajustes de escalas entre *scorings* realizados com dados da rede de autoridades e dos recursos (Robertson e Jones, 1976), dentre outras.

Dessa forma, faz-se necessária a especificação de algoritmos para a RI em sistemas que se baseiem na abordagem de *Folkauthority*. Tais algoritmos devem levar em consideração diversos aspectos em torno da concessão de autoridades, dentre os quais pode-se destacar a popularidade e o peso das *tags* atribuídas às diferentes autoridades e a topologia geral

da cadeia de autoridades, características dinâmicas e que devem ser consideradas pelo algoritmo cada vez que um usuário realiza uma busca.

Nesse sentido, foi realizado um estudo baseado na simulação do processo de concessão de autoridades e na coleta de dados provenientes de um SBF real, denominado *Delicious*[1]. A simulação do processo de concessão de autoridades visa instanciar um modelo que denota a cadeia de autoridades e os recursos categorizados por cada autoridade. Com base na instanciação desse modelo, pretende-se verificar as hipóteses estabelecidas e identificar soluções para as questões relacionadas ao uso do conceito de *Folkauthority* para a RI. A necessidade do emprego da técnica de simulação parte da não existência de um sistema baseado no conceito de *Folkauthority* cujos dados tenham sido gerados a partir da atividade de usuários reais engajados na tarefa de concessão de autoridades (Golbeck e Hendler, 2006).

Sendo assim, neste capítulo é apresentada e discutida a metodologia utilizada para simular a concessão de autoridades cognitivas no contexto introduzido pelo conceito de *Folkauthority*. Além disso, é demonstrada a abordagem para a coleta de dados, na qual define-se a topologia da cadeia de autoridades, a personomia de cada usuário da cadeia e as concessões de autoridade entre os usuários. Apesar da cadeia de autoridades possuir informações que permitem a recuperação de *tags*, autoridades e recursos, os experimentos realizados focam na recuperação de recursos *Web* (*URL's*), os quais são disponibilizados pelos usuários do sistema *Delicious*.

4.1 Um Modelo para o Folkauthority

Existem na literatura diferentes propostas para a modelagem de um SBF. Um modelo comumente utilizado é o modelo baseado em conjuntos. Szomszor et al. (2008) modelam um SBF como sendo uma tupla $\mathbb{F} = (U, T, R, Y)$, na qual o três conjuntos não-vazios, U, T, R representam, respectivamente, os usuários, as *tags* e os recursos, e o produto cartesiano $Y \subseteq U \times T \times R$, o qual denota as *categorizações* no sistema. Hotho et al. (2006) definem formalmente a *personomia* de um usuário $u \in U$, denotada por $\mathbb{P}_u$, como sendo uma restrição a $\mathbb{F}$ tal que $\mathbb{P}_u = (T_u, R_u, I_u)$, sendo $I_u = \{(t, r) \in T \times R : (u, t, r) \in Y\}$, $T_u = \pi_1(I_u)$ e $R_u = \pi_2(I_u)$ e π_i a operação de projeção na i-ésima dimensão da tupla. Outros exemplos de autores que utilizam o modelo baseado em conjuntos incluem (Abel et al., 2008) e (Golder e Huberman, 2006). Begelman et al. (2006) e Halpin et al. (2007) utilizam um modelo baseado em grafos como uma representação alternativa para denotar

[1] http://delicious.com

um SBF. Nesse modelo, $G = (V, E)$ é um hipergrafo tripartite, no qual $V = U \cup T \cup R$ é o conjunto de nós e $E = \{\{u, t, r\} : (u, t, r) \in Y\}$ é o conjunto de hiperarestas.

Os modelos apresentados por esses autores serviram de base para definir formalmente um sistema que utilize o conceito de *Folkauthority*. Nesse contexto, pode-se definir $A \subseteq U$ como o conjunto de autoridades categorizadas, no qual o conjunto A representa as autoridades, enquanto o conjunto U representa os usuários. O conjunto A é um subconjunto impróprio de U, pois as autoridades são os próprios usuários do sistema e inclui o caso em que A é igual a U, isto é, todos os usuários são também autoridades. Essa definição está de acordo com o que é discutido por Pereira (2008), que define que as autoridades em determinada *tag* compõem um subconjunto do conjunto de usuários, conforme ilustrado na Figura 4.1.

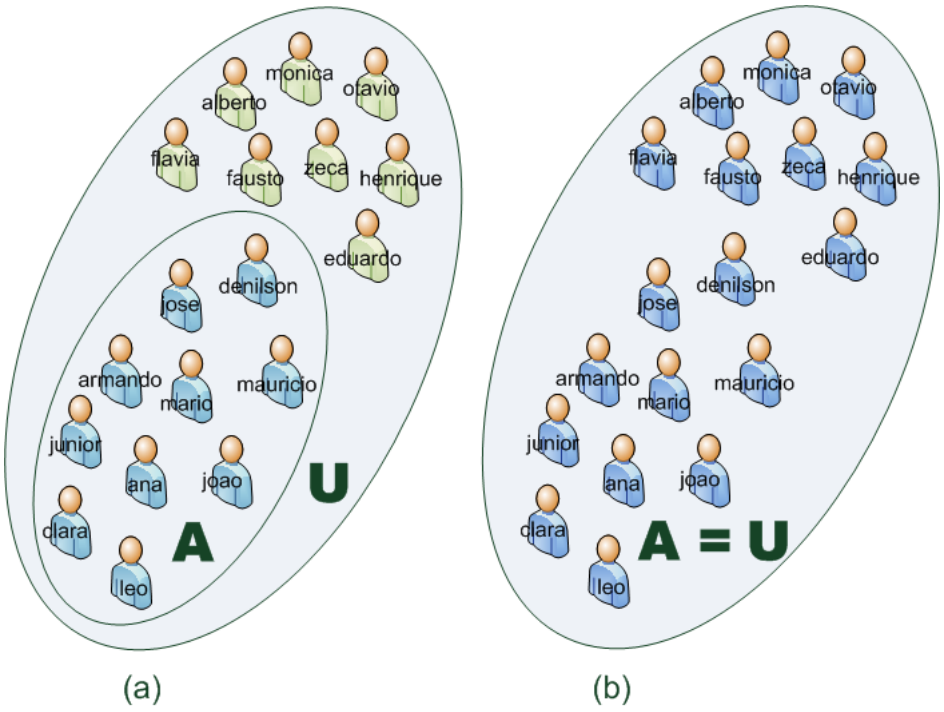

Figura 4.1: Relação entre os conjuntos A e U.

Na figura acima temos que o conjunto A representa as autoridades categorizadas por algum elemento (usuário) do conjunto U. O conjunto $U \setminus A$, cujos elementos são mostrados na Figura 4.1 (a), representam aqueles usuários que não foram ainda categorizados como autoridades. Não obstante, esses usuários podem ter categorizado alguma autoridade no conjunto A. Na parte (b) da figura, mostra-se o caso em que $A = U$, o qual pode representar tanto o fato de que nenhuma autoridade foi concedida e que nenhum usuário

foi adicionado ao sistema (nesses casos, os dois conjuntos estariam vazios), quanto o fato de que foram concedidas autoridades a todos os usuários.

O conjunto de concessões de autoridade é representado pela função $Z : (U \times T \times A) \rightarrow P$, sendo $P = \{1, 2, 3, 4, 5\}$ o conjunto de pesos a serem associados às autoridades. A função $Z(u, t, a) = p$ representa o fato de um usuário u considerar outro usuário a uma autoridade cognitiva no assunto t com peso p, e restringe a aplicação de um único peso à cada autoridade para um assunto específico. Considerando o exemplo demonstrado pela Figura 4.2, verificamos que o usuário denominado de *mario* atribuiu a *tag* `usability` com peso 3 ao usuário denominado de *joao*, sendo, portanto, $Z(mario, usability, joao) = 3$.

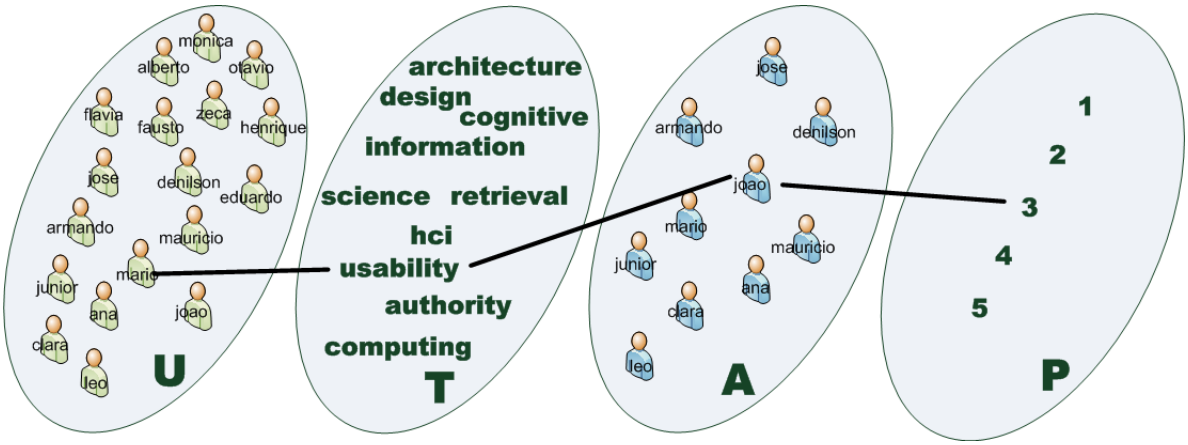

Figura 4.2: A função $Z : (U \times T \times A) \rightarrow P$.

De acordo com a definição de Pereira (2008), a utilização de Folksonomia para a concessão de autoridades cognitivas implica no que pode ser considerado como uma meta-categorização – cada usuário é responsável por categorizar outros usuários, os quais estão acompanhados de suas próprias categorizações (isto é, de suas personomias). As *tags* aplicadas às autoridades são descrições de suas competências e habilidades cognitivas, do ponto de vista de quem as categorizou (Pereira e da Silva, 2008b). Essa situação é ilustrada na Figura 4.3, na qual o usuário denominado de *joao*, considerado autoridade em `usability` por outro usuário denominado de *mario*, está acompanhado de sua personomia, a qual possui os recursos r_1 e r_2 com a *tag* `usability`, o recurso r_3 com a *tag* `hci` e o recurso r_4 com a *tag* `design`. Nesse caso, temos que $\{mario, joao\} \in U$ e que $joao \in A$. Além disso, $\{(usability, r_1), (usability, r_2), (hci, r_3), (design, r_4)\} \in I_{joao}$ e $Z(mario, usability, joao) = 3$ e $Z(mario, hci, joao) = 2$.

Com base nessas definições, foram obtidos dados e elementos para instanciar o modelo de *Folkauthority* apresentado, os quais foram persistidos em um banco de dados relacional.

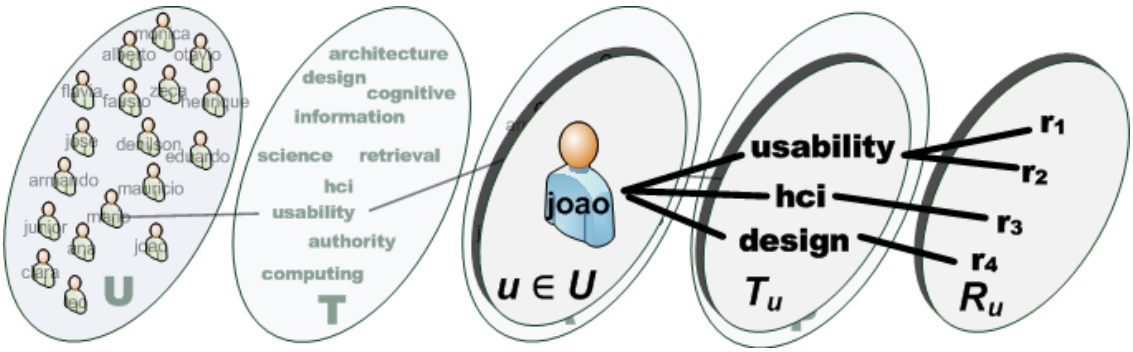

Figura 4.3: A personomia de uma autoridade.

Os dados para a instanciação do modelo foram extraídos de outro SBF, o *Delicious*, cujos recursos são *bookmarks*. Esse sistema possui uma ferramenta chamada de *Network*, cuja utilização gera uma rede social de contatos dentro do sistema. A topologia da cadeia de autoridades foi definida com base nessa rede social de contatos. As justificativas que orientaram as decisões tomadas sobre a escolha das fontes de dados, bem como a metodologia para a obtenção desses dados são descritas na próxima seção.

4.2 Obtenção de Dados

Para que a abordagem para RI desenvolvida possa ser avaliada, é necessário que se tenha disponíveis informações relativas aos usuários, seus recursos e suas *tags*, bem como à cadeia de autoridades, pois deseja-se verificar a recuperação de documentos em um SBF que utilize o conceito de *Folkauthority*. No entanto, obter todas essas informações por meio da atividade de usuários reais demanda um certo período de tempo, bem como um esforço no sentido de prover usuários participantes. Golbeck e Hendler (2006) defendem que simular dados de uma rede social é uma alternativa viável para realização de um estudo/avaliação/validação desses modelos, pois uma rede social formada por usuários reais é difícil de ser obtida, no entanto a maioria dessas redes possuem propriedades topológicas conhecidas, o que facilita a simulação da rede.

Sendo assim, nesta pesquisa a obtenção de dados sobre o processo de categorização dos recursos foi realizada por meio de outro sistema. Uma vez obtidos os dados sobre as categorizações dos usuários, pode-se simular o processo de concessão de autoridades, o qual leva em conta a personomia dos dois usuários que participam desse processo. Por exemplo, a simulação da concessão de autoridade de um usuário A para um usuário B leva em conta que as *tags* com maior frequência de uso na personomia do usuário B são também

as mais prováveis de serem utilizadas na concessão. O resultado dessa simulação é um modelo de *Folkauthority* instanciado, contendo toda a cadeia de autoridades, incluindo o conjunto de recursos e o conjunto de *tags* aplicadas aos recursos por cada usuário. Dessa forma, utilizou-se essa estrutura para comparar uma abordagem para RI que não leve em conta autoridades cognitivas com outra abordagem que leve em conta esse fator. Os testes realizados para a verificação das abordagens, bem como os resultados obtidos, serão discutidos com mais detalhes no próximo capítulo.

A fonte para obtenção de dados para popular o sistema proposto foi o sistema *Delicious*. Esse sistema foi escolhido por causa de suas características e por representar uma base de dados utilizada por diversos outros autores/pesquisadores na área de Folksonomias (Bao et al., 2007; Golder e Huberman, 2006; Kome, 2005; Mathes, 2005; Pereira e da Silva, 2008b; Song et al., 2008; Szomszor et al., 2008). O sistema *Delicious* é um SBF que permite o armazenamento e o compartilhamento de *bookmarks*. Um dos serviços que o sistema *Delicious* oferece é chamado de *Network*, no qual cada usuário do sistema pode manter uma rede de contatos de outros usuários, sendo essa rede assimétrica. Isso significa que um usuário A ao manter o usuário B em sua rede não requer, necessariamente, que o usuário B também mantenha A em sua rede. Esse serviço oferece benefícios, principalmente, no sentido de intermediar e facilitar a visualização das atividades de determinados usuários. Assim, pode-se afirmar que um determinado usuário do sistema provavelmente incluirá em sua rede de contatos aqueles usuários que disponibilizam conteúdos de boa relevância e qualidade, do seu ponto de vista. Nessa ferramenta, um usuário A que inclua outro usuário B em sua rede é chamado de *fã* de B.

Quando um usuário no sistema *Delicious* encontra/conhece outro usuário que costumeiramente disponibiliza informação considerada interessante por esse primeiro usuário, é possível utilizar a ferramenta *Network* a fim de acompanhar as categorizações realizadas por esse último. Outro caso de uso da ferramenta *Network* está em permitir a realização de sugestões de *bookmarks* aos usuários pertencentes a uma determinada rede. Isto é, a ferramenta provê meios simples de compartilhamento de *bookmarks* entre membros de uma mesma rede.

Assim, nessa ferramenta qualquer usuário pode adicionar à sua rede outro usuário que possua *bookmarks* interessantes. No estudo realizado, essa relação é considerada valorosa, pois representa a apreciação da personomia de um usuário por parte de outro usuário. Heuristicamente, pode-se considerar que essa relação representa uma concessão de autoridade cognitiva, uma vez que um usuário demonstra explicitamente seu interesse nos recursos disponibilizados por outro usuário. Não obstante, essa heurística está sujeita a erros, tais como no caso em que um usuário adiciona outro em sua rede por ser um crítico

dos recursos desse último. No entanto, considera-se que esses casos sejam esporádicos e que não interfiram no resultado geral dos experimentos. Deste modo, a topologia da cadeia de autoridades coincidirá com a topologia da rede formada pela utilização da ferramenta *Network*, permitindo definir quais usuários concederam autoridades e quais receberam tal concessão. Além disso, para definir a personomia de cada um desses usuários basta recuperar seus *bookmarks* categorizados.

No entanto, é necessário também definir quais *tags* e quais pesos foram utilizados para descrever as competências cognitivas de cada autoridade. Essa tarefa é realizada por meio da simulação do processo de concessão de autoridades cognitivas, cujas etapas estão descritas na Figura 4.4.

Nessa figura, temos as três etapas do processo de simulação: 1) a definição da topologia da cadeia de autoridades, 2) a definição da personomia dos usuários e, 3) a definição das redes de autoridade. A saída de cada etapa serve de entrada para a etapa posterior. Dessa forma, na primeira etapa é realizada a tarefa de varredura (ou *crawling*) da rede social de contatos formada pela utilização da ferramenta *Network*. Nessa etapa são armazenados números identificadores (*ID's*) de cada usuário da rede, bem como a relação de "quem adicionou quem à sua rede". Por exemplo, quando um usuário A com *ID* 25 adiciona um usuário B com *ID* 26 à sua rede, um par ordenado (25, 26) é armazenado a fim de denotar tal relação. A partir dessas informações, podemos instanciar o conjunto U e o conjunto A, os quais representam os usuários e as autoridades do sistema, respectivamente.

Deste modo, na primeira etapa do processo de simulação do experimento realizado, foram adicionados ao conjunto A uma quantidade de 7210 autoridades e ao conjunto U uma quantidade de 8032 usuários (sendo $|U \setminus A| = 822$), enquanto na segunda etapa foi definida a personomia de cada um desses usuários (incluindo as autoridades). Na definição da personomia de cada um dos usuários foi armazenado um total de 5.873.043 *bookmarks* e 181.505 *tags* diferentes.

A primeira e a segunda etapa foram realizadas de acordo com o Algoritmo 3. Esse algoritmo inicia-se colocando um usuário u e todos os membros da rede desse usuário[2] em uma lista L (a fim de que sejam visitados posteriormente). Também, marca-se u como "visitado" colocando-o na lista V. Por fim, armazena-se todos os *bookmarks* do usuário u juntos de suas respectivas *tags*. Os procedimentos $rede(u)$ e $personomia(u)$ retornam, respectivamente, o conjunto de membros da rede do usuário u e a personomia do usuário u. O algoritmo é repetido para todos os outros usuários colocados na lista L.

[2]A referência ao termo "membros da rede de um usuário u" significa o conjunto de usuários que tenham o usuário u como fã na ferramenta *Network*

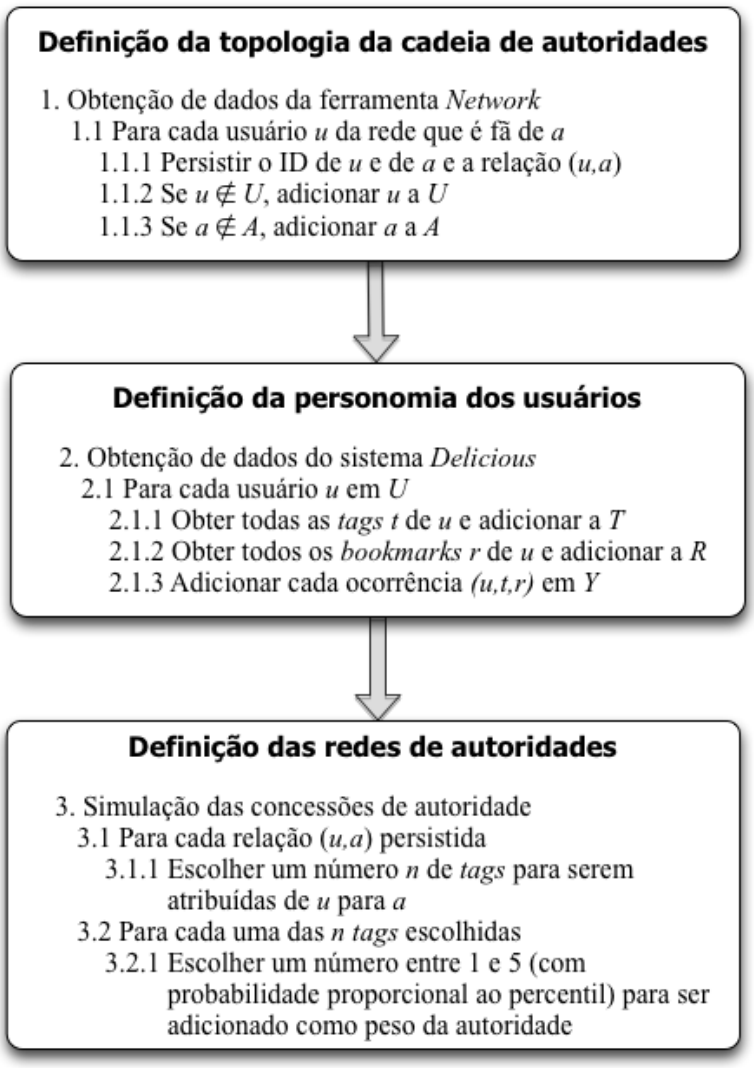

Figura 4.4: Etapas da simulação das concessões de autoridade.

A saída da segunda etapa da simulação é a instanciação dos conjuntos U, T, A, R e Y, além do armazenamento de todas as relações da rede social formada pela utilização da ferramenta *Network*. Essas informações são passadas como entrada para a terceira etapa, na qual simula-se de fato as concessões de autoridade entre os membros dessa rede social.

O processo de simulação das concessões de autoridade foi realizado tendo como premissa que, dada a concessão de autoridade do usuário A para o usuário B, as *tags* que poderiam ser utilizadas para descrever a autoridade cognitiva do usuário B são aquelas presentes na interseção entre o conjunto de *tags* dos dois usuários. Essa heurística é utilizada a fim de modelar o fato de que alguém, para ser considerado autoridade cognitiva

Algoritmo 3. Algoritmo para a definição da topologia da cadeia de autoridades

Entrada: Usuário u
Saída: Definição da topologia da cadeia de autoridades e da personomia dos usuários

$L \leftarrow \{u\}$
$A, U, V \leftarrow \{\}$
para cada $(u \in L) \wedge (u \notin V)$ **faça**
  $A_u \leftarrow rede(u)$
  $A \leftarrow A \cup A_u$
  $U \leftarrow U \cup \{u\}$
  para cada $a \in A_u$ **faça**
    Armazene a relação (u, a)
  fim
  $\mathbb{P}_u \leftarrow personomia(u)$
  $V \leftarrow V \cup \{u\}$
  $L \leftarrow (L \cup A_u)$
fim

por outra pessoa, deveria compartilhar algum vocabulário com essa última. O efeito que essa heurística causa nos resultados da simulação não foi investigado neste trabalho.

A terceira etapa é iniciada escolhendo as *tags* de autoridade a serem atribuídas a algum usuário. Conforme comentado, essas *tags* pertencem à interseção do conjunto de *tags* do usuário que concede e do usuário ao qual é concedida a autoridade. A probabilidade de uma *tag* ser escolhida é baseada na frequência de uso dessa *tag* pelo usuário reconhecido como autoridade. Além disso, outras três questões são consideradas:

- Se a interseção entre o conjunto de *tags* dos dois usuários é vazia, então nenhuma concessão de autoridade deverá ocorrer;
- O número de *tags* a serem utilizadas na concessão de autoridade deve ser próximo da média de uso de *tags* nos recursos[3];
- O nível a ser atribuído a uma *tag* utilizada na concessão de autoridade deve ser proporcional à frequência de utilização dessa *tag* pela autoridade.

Considerando todas essas questões, a terceira etapa da simulação do processo de concessão de autoridades foi definida conforme ilustrado na Figura 4.5, a qual mostra a sequência evidenciando os passos e o fluxo dessa etapa da simulação.

[3]Conforme discutido por Golder e Huberman (2006), no sistema *Delicious* cada usuário utiliza uma média de três *tags* em cada categorização.

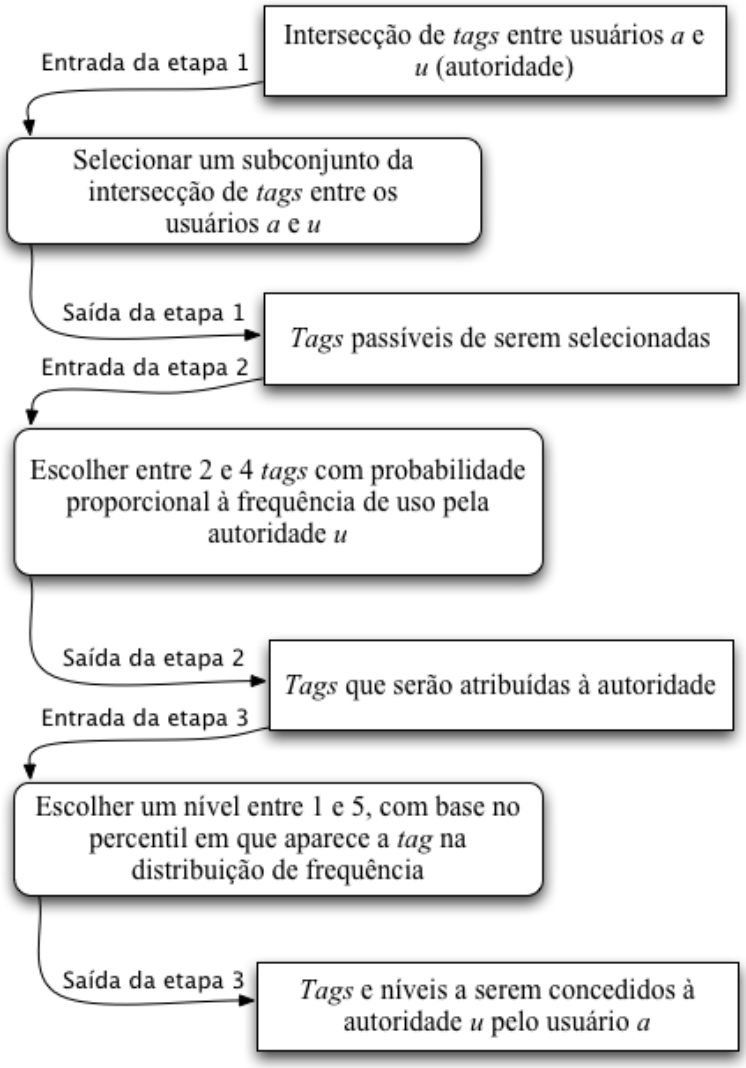

Figura 4.5: Etapas da simulação das concessões de autoridade.

Nessa figura, considera-se a concessão de autoridade de um usuário a para um usuário u. Inicialmente, um subconjunto da interseção de *tags* desses usuários é selecionado, as quais são consideradas passíveis de serem atribuídas à autoridade. Esse subconjunto é determinado com base na noção de *long tail* (Halpin et al., 2007; Sen et al., 2006). O termo *long tail* refere-se ao efeito de "cauda longa" mostrado na Figura 4.6, a qual ilustra o histograma da distribuição de frequência de uso de *tags* de um usuário hipotético, gerado a partir da Tabela 4.1. Essa distribuição é tipicamente chamada de "curva de lei de potência" e descreve os níveis abruptos de decaimento na utilização de determinadas *tags*. Isto é, há uma pequena parcela das *tags* que ocorrem muitas vezes, junta de uma grande

parcela das *tags* que ocorrem poucas vezes. Essa distribuição está também relacionada com o *Princípio de Pareto*, que descreve o comportamento de sistemas nos quais 80% dos efeitos são gerados por 20% das causas, e nos quais 80% das causas geram 20% dos efeitos (Golder e Huberman, 2006; Halpin et al., 2007; Sen et al., 2006).

Tag	**Frequência de utilização**
`information`	10
`index`	9
`retrieval`	8
`folksonomy`	4
`authority`	4
`research`	2
`web`	2
`quality`	1
`ranking`	1
`relevance`	1
`lucene`	1
`text`	1
Total	27

Tabela 4.1: Distribuição de uso de *tags* de um usuário típico.

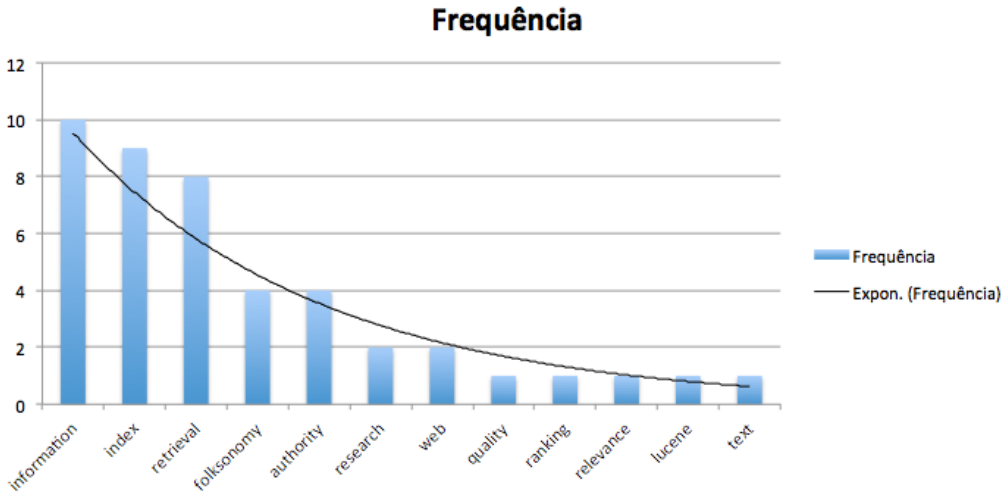

Figura 4.6: *Long tail* de *tags*.

No processo de escolha das *tags* candidatas a serem adicionadas a uma autoridade desconsiderou-se 75% das *tags* da autoridade – justamente aquelas cuja frequência de utilização são consideradas pequenas quando comparadas com a frequência de uso de

outras *tags* mais frequentes. No exemplo da Tabela 4.1, tem-se 12 *tags*, sendo que 25% representam três *tags*. Essas três *tags* são `information`, `index` e `retrieval`, já que as demais pertencem ao que é considerado o *long tail*. Com base nesse cálculo, o conjunto de *tags* {`information`, `index`, `retrieval`} deverá passar por uma operação de intersecção com o conjunto de *tags* do usuário a. O resultado dessa operação é o conjunto de *tags* passíveis de serem atribuídas ao usuário u como autoridade.

Deste modo, cada *tag* é escolhida com base na frequência de uso pelo usuário B. Por exemplo, suponha que as *tags* passíveis de serem atribuídas à autoridade, com suas respectivas frequências de uso, são: `information` – 10 utilizações, `index` – 9 utilizações e `retrieval` – 8 utilizações. Considere também que a probabilidade $p(t)$ de uma *tag* t ser escolhida é igual $freq(t)/n$, sendo n a soma da frequência de uso das *tags* consideradas e $freq(t)$ a frequência de uso da *tag* t pelo usuário B. Nesse exemplo, teremos $p($`information`$) = 10/27 = 0.37$, $p($`folksonomy`$) = 9/27 = 0.34$ e $p($`authority`$) = 8/27 = 0.29$. Essa idéia é descrita no Algoritmo 4, no qual cada *tag* passível de ser aplicada à autoridade é escolhida com base em uma probabilidade previamente calculada.

Algoritmo 4. Algoritmo para definição das *tags* utilizadas nas concessões de autoridade

Entrada: Lista L com as *tags* passíveis de serem adicionadas à autoridade
Saída: *Tag* selecionada para ser adicionada à autoridade

```
n ← numAleatorio(0.0 − 1.0)
i ← 0.0
para cada t ∈ L faça
    i ← i + p[t]
    se i > n então
        retorna t
    fim
fim
```

Na listagem Algoritmo 4, uma lista L contendo as *tags* candidatas a serem atribuídas à autoridade é tomada como entrada. Para cada elemento t nessa lista, há um elemento correspondente no vetor $p[t]$, cujo conteúdo refere-se à probabilidade desse elemento ser escolhido. O método de escolha do elemento é baseado no método da roleta, no qual um número aleatório entre 0.0 e 1.0, chamado de n, é gerado. Após isso, itera-se sobre os elementos do vetor $p[\,]$, acumulando-se na variável i a probabilidade de cada elemento do vetor ser escolhido. Quando essa variável atinge ou ultrapassa o valor do número aleatório n gerado, o último elemento cuja probabilidade foi adicionada à variável i é escolhido. Esse procedimento é repetido entre 2 e 4 vezes, com cada repetição retirando o elemento

escolhido t da lista passada como entrada. O número de vezes que o procedimento repete é definido aleatoriamente.

Por fim, deve-se decidir sobre os níveis de autoridade que deverão acompanhar as *tags* escolhidas. Essa decisão é baseada no percentil no qual a *tag* selecionada está localizada na distribuição de frequência de uso das *tags* da autoridade. Um exemplo pode ser descrito voltando-se à Figura 4.6. No histograma mostrado nesta figura, podemos particionar a distribuição em cinco grupos, denominados *percentis*. O primeiro percentil compreende 100-80% das *tags* mais utilizadas, enquanto o segundo percentil compreende entre 80-60% das *tags* mais utilizadas. O terceiro percentil compreende entre 60-40%, e assim por diante. Dessa forma, definiu-se que, dentre as *tags* passíveis de serem atribuídas à autoridade (isto é, as *tags* fora da cauda longa), uma que fosse escolhida e que pertencesse ao primeiro percentil receberia um peso igual a 5, enquanto uma que pertencesse ao segundo percentil receberia um peso igual a 4. Uma *tag* que pertencesse ao terceiro percentil receberia peso 3, e assim por adiante.

Em cada um desses passos, foram consideradas questões que atendessem as necessidades apresentadas sobre a simulação. Por exemplo, o número de *tags* a serem atribuídas às autoridades está compreendido entre 2 e 4. Essa escolha é baseada no fato de que a média de *tags* utilizadas em uma categorização é 3, conforme demonstrado em (Golder e Huberman, 2006; Wang e Davison, 2008). Além disso, definiu-se os termos da concessão de autoridade com base na intersecção do vocabulário dos usuários que participam dessa relação. Dessa forma, pretende-se que a simulação reflita o fato de que esses usuários provavelmente compartilhem algum vocabulário, pois do contrário dificilmente participariam de uma relação de concessão de autoridade cognitiva. Por fim, as *tags* e os pesos das *tags* atribuídas às autoridades são definidas com base na quantidade de vezes que a autoridade as utilizou.

4.3 Um SRI que utiliza o conceito de Folkauthority

Após definidos o modelo para representar um sistema baseado no conceito de *Folkauthority* e os dados necessários para instanciar esse modelo, procedeu-se com a implementação de um sistema que fornecesse uma interface de busca sobre esses dados. O sistema desenvolvido foi chamado de *AuthoritySearch*. Esse sistema possui os componentes básicos de um SRI (Baeza-Yates e Ribeiro-Neto, 1999; Manning et al., 2009), os quais estão na Figura 4.7. O componente central dessa figura é o índice, o qual é responsável por armazenar de maneira eficiente os documentos do SRI. Além disso, pode-se notar a distinção de duas partes na figura, consoantes a dois momentos distintos na RI: os

componentes relacionados com o *tempo de consulta* e os componentes relacionados com o *tempo de indexação*. Em tempo de indexação, os documentos (ainda em uma forma não estruturada) são coletados e analisados, realizando um processamento pré-indexação. Por fim, as informações são armazenadas no índice invertido, de forma que uma consulta possa ser realizada eficientemente. Em tempo de consulta, um usuário utiliza a interface de busca para realizar uma consulta no sistema. O ambiente de execução da consulta calcula a similaridade entre a consulta e os documentos presentes no índice, oferecendo ao usuário um conjunto ordenado de resultados de busca.

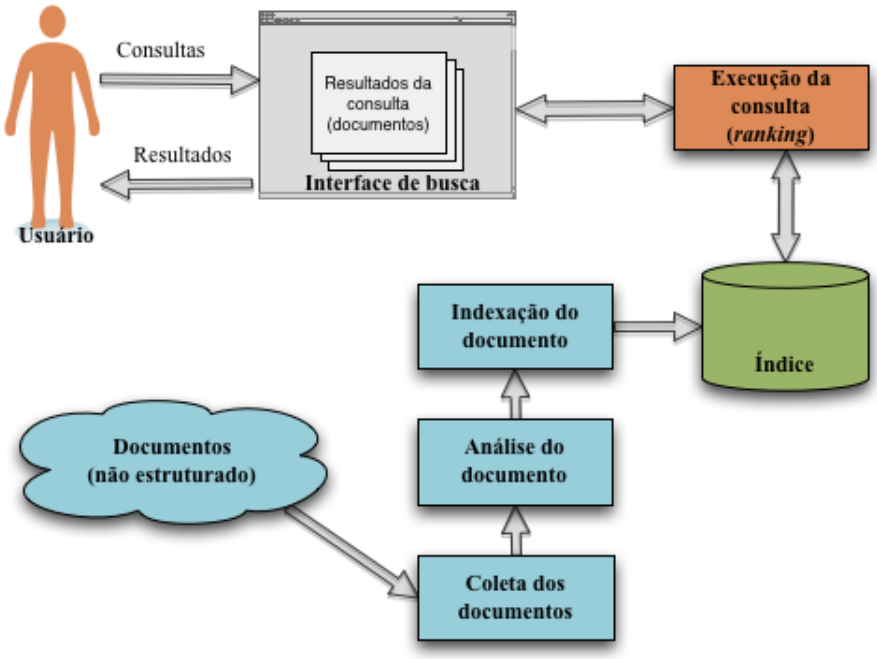

Figura 4.7: Elementos de um SRI.

Os componentes de *coleta do documento*, *análise do documento* e *indexação do documento* realizam um processamento nos documentos, anterior à indexação. Esse processamento geralmente inclui a coleta de metadados (Angeletou et al., 2007), o particionamento do texto em termos de índice (Zobel e Moffat, 2006) e a utilização de técnicas como a remoção de *stop words* e de *stemming* (Frakes e Baeza-Yates, 1992).

A *interface de busca* está relacionada com os componentes de um SRI com os quais o usuário pode interagir. A qualidade do projeto desses componentes influenciam diretamente a eficiência da tarefa de RI (Jin e Fine, 1996). Por meio dessa interface o usuário pode expressar sua necessidade de informação (consulta) e, possivelmente, satisfazê-la a partir da avaliação de uma coleção de documentos que atendem à uma consulta (resultado

de uma busca). Existem diversos autores que discutem questões de projeto para interfaces de busca. De modo geral, as questões que orientam o projeto e o desenvolvimento dessas interfaces apoiam-se em metas de usabilidade e de experiência do usuário (Hearst, 2009; Jin e Fine, 1996; Rogers et al., 2002).

Por fim, o componente de *execução da consulta* realiza uma série de tarefas. Uma delas é a tarefa de processamento da consulta, na qual realiza-se um *parsing* da consulta, extrai-se todos os termos avaliados de uma expressão booleana e, opcionalmente, adiciona-se um peso diferenciado a alguns termos. Outra tarefa é a busca no índice, na qual encontra-se a coleção de documentos que atendem à consulta. Além disso, esse componente é responsável por realizar um cálculo da similaridade entre a consulta e os documentos do índice, bem como ordenar os documentos com base no valor dessa similaridade. No sistema *AuthoritySearch*, os modelos teóricos utilizados para a implementação desse componente foram o modelo booleano e o modelo de espaço vetorial (Baeza-Yates e Ribeiro-Neto, 1999; Manning et al., 2009).

Para a tarefa de RI no sistema *AuthoritySearch*, há a necessidade de que o cálculo do *ranking* de um documento reflita a sua importância com base nas autoridades que categorizaram esse documento. Tal requisito levanta algumas questões de projeto para uma abordagem de RI em um sistema que utiliza o conceito de *Folkauthority*. Essas questões são descritas a seguir.

- É necessária a determinação de um modelo apropriado (ou de uma combinação de modelos apropriados) para representar a importância das categorizações realizadas pelas autoridades. Deve também ser descrito um algoritmo baseado nesse modelo que seja capaz de realizar um cálculo de semelhança entre uma consulta e os documentos, levando em consideração as autoridades que categorizaram os documentos. Além disso, esse modelo deve permitir a realização de consultas no sistema de forma eficiente.

- O cálculo da contribuição de cada autoridade para a importância de um documento, dada uma consulta, deve ser baseado nas *tags* que essa autoridade utilizou para categorizar esse documento. Por exemplo, uma autoridade à qual tenha sido atribuída uma *tag* `information` deve possuir documentos em sua personomia categorizados com essa *tag*. Esses documentos, conforme discutido, terão o valor de seus *rankings* diferenciados quando uma consulta contendo o termo `information` for passada ao sistema. Dessa forma, é necessária uma representação que possa distinguir a importância de cada *termo* utilizado por uma autoridade para descrever um

documento a fim de que, em tempo de busca, essa distinção de importância seja refletida no *ranking* de cada documento.

- Deve-se decidir sobre quais aspectos da cadeia de autoridades devem ser considerados no cálculo do *ranking* de cada documento. Conforme apresentado em (Pereira e da Silva, 2008b), a cadeia de autoridades fornece informações sobre a popularidade e o peso de cada autoridade. Portanto, é necessária a definição de um modelo que sintetize todos esses aspectos, de forma que o cálculo da importância de uma autoridade para determinada *tag* em um documento possa refletir todos esses (ou boa parte desses) aspectos.

- Há uma questão sobre *quando* realizar o cálculo das informações relacionadas a cadeia de autoridades. Uma alternativa é calculá-las em tempo de indexação, de forma que a importância de cada autoridade para cada termo dos documentos seja calculada *offline* e então atribuída aos documentos nos índices, na forma de um *boost* nos termos dos documentos. Quando uma consulta ocorre, o valor desse *boost* já está previamente calculado. Essa abordagem possui a desvantagem de necessitar uma atualização do valor do *boost* no índice quando uma autoridade, uma *tag* ou um documento é adicionado/removido do sistema. No entanto, essa abordagem oferece vantagens quanto a eficiência de tempo da busca. Outra alternativa é calcular as informações relacionadas a cadeia de autoridades em tempo de consulta. Nessa abordagem, a importância de cada autoridade para cada termo nos documentos deve ser estipulada a cada consulta realizada no sistema. Essa abordagem elimina a necessidade da atualização constante dos valores de *boost* no índice, no entanto pode comprometer a eficiência de tempo da busca.

A Figura 4.8 apresenta a arquitetura do sistema *AuthoritySearch*, a qual foi derivada das decisões de projeto que serão apresentadas. Essa arquitetura possui um *módulo de interface com o usuário*, por meio do qual um usuário pode realizar uma consulta no sistema e receber como retorno uma lista ordenada dos documentos que atendem à consulta. Há ainda um *módulo de cálculo da similaridade entre uma consulta e os documentos*, o qual é responsável por calcular o *ranking* de cada documento que atende a uma consulta, e um *módulo de indexação*, o qual é responsável por armazenar as informações de índice em estruturas de dados necessárias para um acesso eficiente às informações. Esse módulo utiliza serviços do *módulo de armazenamento das personomias*, o qual fornece informações sobre quais usuários categorizaram quais documentos e com quais *tags*. Finalmente, tem-se o *módulo de cálculo de ranking das autoridades*, o qual é

responsável por calcular a importância de cada autoridade em um assunto. Esse módulo utiliza informações da rede de autoridades para realizar tal tarefa e se comunica com o módulo de indexação, pois precisa informar quais os valores dos *boosts* de cada termo utilizado por uma autoridade.

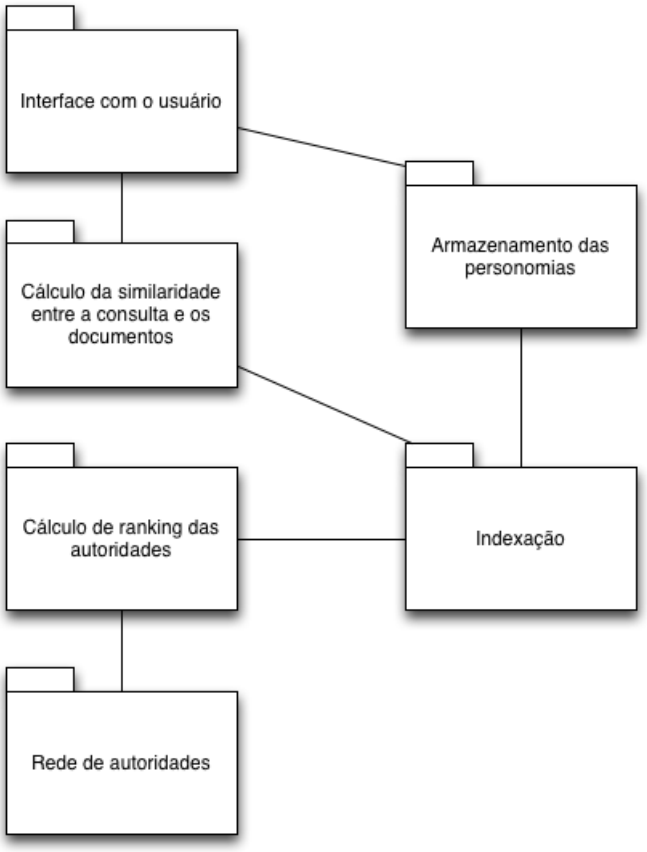

Figura 4.8: Arquitetura do sistema *AuthoritySearch*.

Nas sub-seções seguintes, serão descritos os processos de busca, *ranking* e indexação de documentos no sistema *AuthoritySearch*. Esses processos foram elaborados de acordo com as decisões que foram tomadas com base nas questões de projeto descritas.

4.3.1 Busca e Ranking no Sistema

Em um SRI a busca se inicia a partir da necessidade de informação de um usuário (Manning et al., 2009). No entanto, do ponto de vista do funcionamento do sistema, a busca é o processo de procura por termos no índice invertido, avaliando os documentos que possuem determinado termo. Um usuário que possua uma necessidade de informação,

por sua vez, realiza uma consulta no sistema e obtém uma série de informações que, supostamente, satisfazem sua necessidade. Essa "suposição" é realizada por meio de cálculos que tentam medir uma similaridade entre a consulta que o usuário realizou e cada documento presente no índice. Os documentos que são avaliados como sendo mais semelhantes a consulta realizada pelo usuário são mostrados no topo das listas de resultados de busca (isto é, na lista que contém os documentos presentes no índice que possuem os termos utilizados na consulta). Esse processo de ordenação dos documentos, dada uma consulta, é geralmente chamado de *ranking*. O valor que reflete a similaridade entre uma consulta e um documento é também chamado de *score*. Essa seção apresenta como o *ranking* de documentos acontece no sistema *AuthoritySearch*. Na implementação desse sistema, o suporte às tarefas de análise dos documentos, da execução da busca e da indexação dos documentos foi realizado pelo *framework Lucene*[4].

A partir da interface mostrada na Figura **??** o usuário é capaz de utilizar operadores booleanos para compor sua consulta (McCandless et al., 2010). Quando a consulta é passada ao sistema, um *parser* realiza uma avaliação dessa consulta, transformando-a em uma representação que possa encontrar documentos no índice que atendam à consulta de forma eficiente. Esse *parser* avalia possíveis utilizações de termos booleanos na consulta, bem como possibilita a utilização de aspas duplas para a utilização de frases exatas e a utilização do operador * para indicar uma parte de um termo qualquer em uma consulta. Por exemplo, uma consulta do tipo "`information retrieval`" `and index`* deverá retornar aqueles documentos contendo a frase "`information retrieval`" *e também* contendo termos que iniciam com `index`, tais como `indexing` e `indexes`. Essa consulta deverá ser representada de forma a expressar essas condições (McCandless et al., 2010).

Quando a consulta é preparada e representada na forma apropriada, o processo de *ranking* das informações se inicia. Para cada documento $\vec{d_j}$ que atenda à uma consulta $\vec{q}$ calcula-se um valor $s_{j,q}$ chamado de *score*, o qual representa a similaridade entre um documento d_j e a consulta $\vec{q}$. Esse valor é calculado tendo como base os termos t_i utilizados na consulta $\vec{q}$. Um resultado de busca é constituído desses documentos ordenados pelos seus respectivos valores de *score*. Para calcular o *score* de cada documento j dada uma consulta q utiliza-se a Equação 4.1.

$$s_{j,q} = \sum_{t_i \in q} (tf_{i,j} \times idf_i) \times boost_{i,j} \tag{4.1}$$

O cálculo dos valores de $tf_{i,j}$ e idf_i foram detalhadamente explicados no Capítulo 2. O valor de $tf_{i,j}$ representa a importância de um termo t_i presente em um documento

[4]http://lucene.apache.org/

d_j. O valor de idf_i refere-se à importância de um determinado termo t_i para distinguir documentos no índice. O esquema de pesos *tf-idf* é tradicionalmente utilizado em SRIs. O valor de $boost_{i,j}$ é calculado em tempo de indexação, com base nas autoridades que categorizaram o documento d_j. Esse incremento é um peso para um termo t_i presente no documento d_j. O cálculo desse fator será explicado na sub-seção seguinte, na qual será tratada a indexação no sistema *AuthoritySearch*.

Após serem calculados o valor de $s_{j,q}$ para cada documento d_j, esses documentos são ordenados de acordo com o referido valor e então apresentados ao usuário. A Figura 4.10 mostra a interface do sistema *AuthoritySearch* pela qual os resultados de uma busca são apresentados.

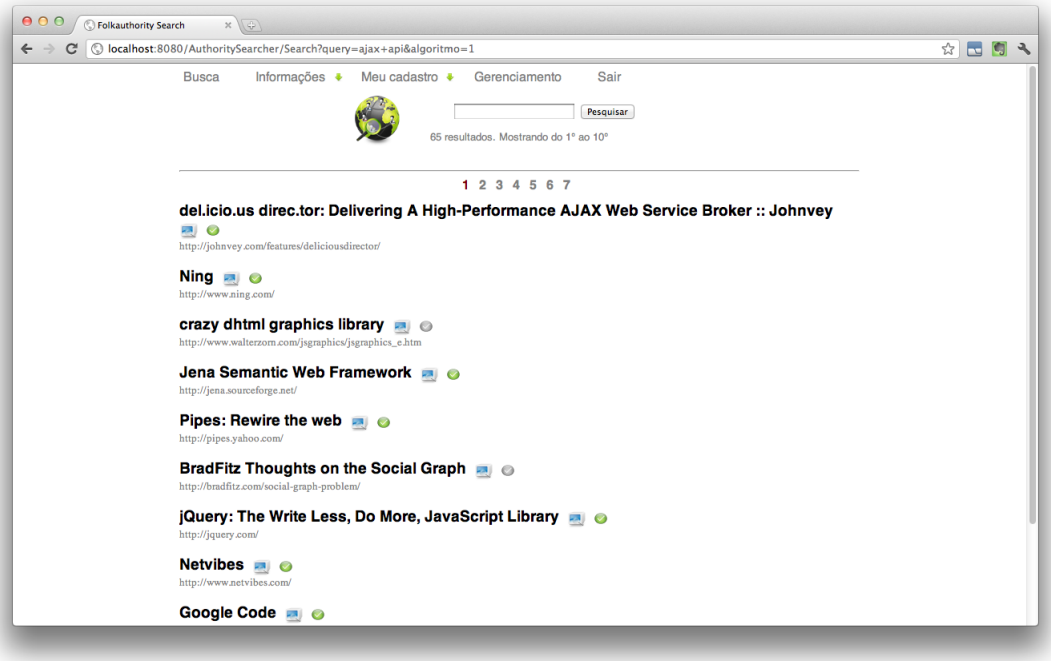

Figura 4.9: Resultados de uma busca no sistema *AuthoritySearch*.

O cálculo do valor de $s_{j,q}$ se assemelha com o cálculo realizado no Algoritmo 2, com a diferença de que o valor de *boost* em cada termo de um documento é levado em consideração na abordagem proposta. Esse incremento representa no *ranking* a importância do fato de uma autoridade ter categorizado o documento.

Outra questão a ser destacada é a possibilidade de reordenação dos n primeiros resultados de uma busca realizada utilizando-se o *score* calculado por meio de $s_{j,q}$. Na Equação 4.1 é possível verificar duas componentes distintas, uma relacionada ao cálculo de $tf_{i,j} \times idf_i$ e outra relacionada ao cálculo de $boost_{i,j}$. Assim, é possível uma reordenação dos

n primeiros resultados com base nessas componentes. Um exemplo pode ser descrito na Tabela 4.2. A parte superior da tabela mostra um resultado de busca hipotético, no qual os cinco primeiros documentos foram recuperados pelo esquema AR. Nessa tabela temos as informações de identificação dos documentos, do cálculo do *score* por meio da equação de $s_{j,q}$ e do cálculo da componente $\sum_{t_i \in q}(tf_{i,j} \times idf_i)$ da equação. Temos também, na parte inferior da Tabela 4.2, a reordenação desses cinco resultados de busca com base no valor da componente $\sum_{t_i \in q}(tf_{i,j} \times idf_i)$. Nessa reordenação, o esquema de AR realiza a tarefa de selecionar os n documentos mais semelhantes com a consulta, enquanto o esquema *tf-idf* realiza a tarefa de ordenar esses n documentos com base no esquema *tf-idf*. Isto é, o esquema de AR serve como um "filtro" dos documentos no índice, enquanto o esquema de *tf-idf* realiza a ordenação somente nesses n elementos recuperados com o AR (diferente de um esquema que calculasse o valor de *tf-idf* para ordenar todos os documentos do índice).

Ordem	**Documento**	***Score*** $(s_{i,j})$	**Componente *tf-idf*** $(\sum_{t_i \in q} tf_{i,j} \times idf_i)$
1	Doc3	3,14	2,76
2	Doc1	2,78	1,64
3	Doc2	2,14	1,54
4	Doc5	1,98	1,81
5	Doc4	1,17	1,05

⇓ Reordenação

Ordem	**Documento**	***Score*** $(s_{i,j})$	**Componente *tf-idf*** $(\sum_{t_i \in q} tf_{i,j} \times idf_i)$
1	Doc3	3,14	2,76
2	Doc5	1,98	1,81
3	Doc1	2,78	1,64
4	Doc2	2,14	1,54
5	Doc4	1,17	1,05

Tabela 4.2: Reordenação com base no cálculo de TF-IDF dos resultados de uma busca com AuthorityRank

No decorrer deste trabalho foi avaliada a utilização do modelo associativo para recuperar informação com a abordagem de *Folkauthority* (Cogo e da Silva, 2010), o qual mostrou-se um modelo adequado para tal tarefa. No entanto, os algoritmos apresentados neste trabalho tiveram suas implementações baseadas no modelo de espaço vetorial, pelo fato dessas implementações terem apresentado bons resultados em termos de eficiência, uma vez que foram utilizadas ferramentas bem estabelecidas para os cálculos de *score* e as operações no índice invertido. Ainda assim, considera-se que o modelo associativo é mais adequado para representar as relações entre documentos, *tags* e autoridades, no sentido

de não requerer nenhuma representação além dessas três entidades e suas relações para calcular um *ranking* dos documentos ou das autoridades. Isto é, o modelo associativo possui uma *tradução mais adequada*[5] com relação às condições impostas à cadeia de autoridades e aos recursos em um sistema que utilize a abordagem de *Folkauthority*.

4.3.2 Indexação no sistema

No sistema *AuthoritySearch* os documentos são indexados levando em conta *quem* os disponibilizou. Por exemplo, considere um documento contendo a *tag* `indexing` e que tenha sido disponibilizado por uma autoridade em `indexing`. De acordo com as discussões apresentadas sobre a RI utilizando o conceito de *Folkauthority*, uma busca utilizando o termo `indexing` na consulta deve priorizar (aumentando o valor do *score*) aqueles documentos categorizados com o referido termo. Sendo assim, foi adotada uma abordagem na qual um fator de *boost* é dado aos termos que aparecem nos documentos categorizados por uma autoridade nesse termo. Esse fator de *boost* é determinado com base na popularidade dessa autoridade na cadeia de autoridades, considerando os pesos e as *tags* atribuídas à ela.

Por exemplo, na Figura 4.10 é ilustrada uma situação na qual foi concedida autoridade a um usuário com as *tags* `information` e `quality` com os pesos 3 e 2, respectivamente. Além disso, esse usuário categorizou os documentos d_1 e d_2 com a *tag* `information`, os documentos d_3 e d_4 com a *tag* `quality` e o documento d_5 com as *tags* `information` e `quality`. Dessa forma, o fator de *boost* a ser calculado para o termo `information` nos documentos d_1 e d_2 será baseado no peso que foi dado à *tag* `information` atribuída à autoridade que categorizou os recursos. Nesse caso, o *boost* a ser dado ao termo `information` nos documentos d_1 e d_2 será maior do que o *boost* a ser dado ao termo `quality` nos documentos d_3 e d_4, já que o peso dado à autoridade no termo `information` é maior do que o peso dado no termo `quality`.

O documento d_5 da Figura 4.10 acumula os *boosts* provenientes dos termos `information` e `quality`, uma vez que o documento foi categorizado com os dois termos e pela Equação 4.1 o *boost* de cada termo presente na consulta e no documento é somado ao *score* final do documento. Além disso, se nesse exemplo houvesse outros usuários na cadeia de autoridades, atribuindo a *tag* `quality` à autoridade, o *boost* do termo `quality`

[5]Goldbarg e Luna (2005) afirmam que "*Um bom modelo exige uma conveniente tradução contextual. Uma boa tradução contextual pode ser expressa através de um correto isomorfismo entre o fenômeno e seu modelo.*" No caso do modelo associativo para a cadeia de autoridades, a representação (grafo ponderado orientado) de fato apresenta uma grande semelhança com o objeto sendo modelado (cadeia de autoridades).

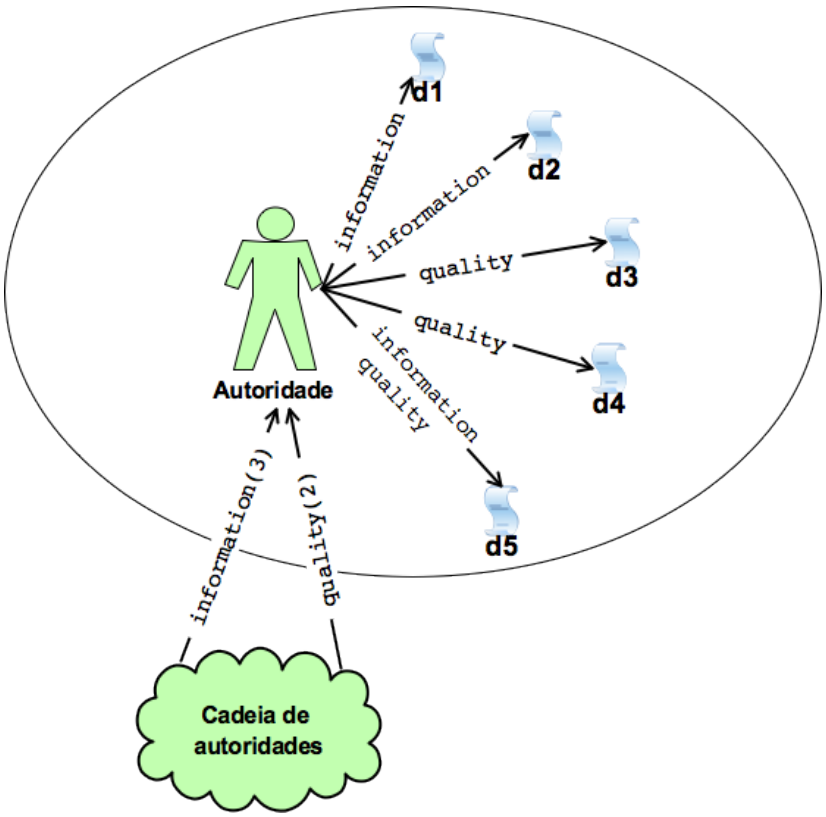

Figura 4.10: Processo de atribuição de *boost* aos termos em um documento.

nos documentos poderia ser ainda maior – talvez maior do que o *boost* introduzido pelo termo `information`, ainda que esse termo possua um peso maior atribuído à autoridade do que o termo `quality`. Isso ocorrerá porque considera-se que a popularidade de uma autoridade também deve ser considerada no cálculo do *ranking* dos documentos (Pereira e da Silva, 2008b) e, a medida que outros usuários também atribuíssem a *tag* `quality` à autoridade, maior seria o acúmulo de *boost* para esse termo nos documentos pertencentes à autoridade.

Outra característica que foi considerada neste trabalho provém do fato de uma autoridade em um assunto poder ser categorizada por outra autoridade nesse mesmo assunto. A Figura 4.11 ilustra essa ideia. Pode-se ver nessa figura um encadeamento de concessão de autoridades. O usuário rotulado por A considera o usuário rotulado por B autoridade no assunto `information` com peso 3. O usuário rotulado por B, por sua vez, considera o usuário rotulado por C autoridade no assunto `information` com peso 2. Da mesma forma, o usuário rotulado por D também é considerado autoridade no assunto `information`, com peso 3. Levando em conta as premissas estabelecidas nas discussões sobre o conceito de autoridade cognitiva e *Folkauthority*, pode-se afirmar que a contribuição para o fator de

boost proveniente do usuário C pode ser maior do que a contribuição proveniente do usuário D, mesmo tendo esse último um peso maior associado a concessão de autoridade cognitiva. Isso ocorrerá porque a autoridade C é categorizada por outra autoridade no mesmo assunto. Chama-se esse fenômeno de *propagação de autoridade* (Pereira e da Silva, 2008b) e, conforme comentado, essa característica foi expressa no *ranking* dos documentos. É importante também perceber que o grafo que representa a cadeia de autoridades pode gerar ciclos e relações de transitividade entre as concessões de autoridade, sendo que tais questões são tratadas a partir de um cálculo iterativo do valor de *PageRank* à Priori, o qual itera sobre todas as arestas do grafo restringido por uma *tag*.

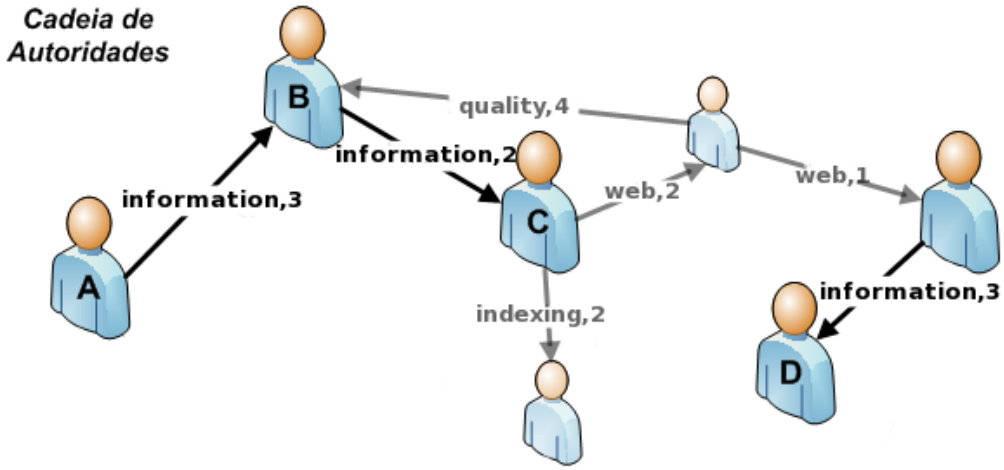

Figura 4.11: Propagação da contribuição dos *boosts* das autoridades.

Para que a característica de propagação de autoridade esteja presente no *ranking*, é necessário um esquema de cálculo de peso das autoridades na rede que considere esse fenômeno. Sendo assim, optou-se por realizar o cálculo do *PageRank* à Priori das autoridades (White e Smyth, 2003), a fim de representar o valor da contribuição de cada autoridade para o *boost* de um termo em um documento. Esse cálculo é mostrado na Equação 4.2 (essa equação é uma instância da Equação 2.5, Capítulo 2, Subseção 2.2.2), na qual temos que pr_i é o valor do *PageRank* à Priori para uma autoridade $a_i \in A$. Além disso, $u_j \in U$ refere-se ao usuário que categorizou a autoridade a_i, $t_k \in T$ e $p_m \in P$ referem-se, respectivamente, à *tag* e ao peso atribuídos a autoridade a_i pelo usuário u_j (p_m refere-se ao valor de $p(v|u)$ na Equação 2.5). O componente pr_j é o valor do *PageRank* à Priori do usuário que categorizou a autoridade a_i enquanto o_j é o número de autoridades categorizadas pelo usuário u_j (a quantidade de arestas de saída do nó u_j na rede de autoridades). Dessa forma, o valor do *PageRank* à Priori para uma autoridade a_i em um

termo t_k é baseada na soma do *PageRank* das autoridades que categorizaram a autoridade a_i. É importante também ressaltar que o grafo representado pela cadeia de autoridades pode, em certas ocasiões, apresentar ciclos e arestas bidirecionais, os quais não impedem o cálculo de *PageRank* à Priori, uma vez que o algoritmo itera sobre o *conjunto* de arestas, dessa forma não apresentando repetição de arestas.

$$pr_i = \sum_{Z:U_j \times T_k \times A_i \rightarrow p_m} p_m \times (pr_j / o_j) \tag{4.2}$$

O valor do *PageRank* à Priori denota a importância de uma determinada autoridade na cadeia de autoridades (Page et al., 1998; White e Smyth, 2003) e foi calculado neste trabalho com o auxílio da ferramenta *JUNG – Java Universal Network/Graph Framework*[6], cuja implementação é baseada em White e Smyth (2003). Esse valor foi utilizado como fator de *boost* para um termo em um documento. A decisão de calcular o fator de *boost* em tempo de indexação, ao invés de calculá-lo em tempo de busca, foi tomada tendo como base o fato de ter sido possível alcançar um bom desempenho na execução de uma consulta com essa abordagem (pois o valor do *PageRank* não precisará ser calculado no momento da consulta). Além disso, no sistema considerado para avaliação da abordagem de RI utilizando o conceito de *Folkauthority*, o índice não se modifica, não sendo necessário recalcular os valores do *PageRank* a cada recurso indexado. Esses fatores levaram a decisão de utilizar a abordagem de cálculo dos *boosts* em tempo de indexação.

O processo de indexação inicia-se com a extração de termos (*tags*) dos documentos. Para cada *tag* do documento a ser indexado, determina-se a cadeia de autoridades para esse termo. Um exemplo dessa situação pode ser ilustrado na Figura 4.11. Nessa figura pode-se ver uma cadeia de autoridades com base na *tag* `information`. Uma cadeia de autoridades como essa é determinada para cada termo de cada documento, empregando o valor do *PageRank* à Priori da autoridade que categorizou o documento em questão como valor de *boost* do termo no documento.

Com base nessas definições, a indexação no sistema *AuthoritySearch* foi definida conforme os passos mostrados no Algoritmo 5. Esse algoritmo recebe como entrada uma lista D contendo os documentos a serem indexados. Para cada documento, calcula-se o *boost* de seus respectivos termos com base no *PageRank* das autoridades que categorizaram esses documentos. O procedimento $termos(d_j)$ retorna uma lista com os termos presentes no documento d_j, a qual é atribuída a T_j. A lista R_k armazena as informações sobre as concessões de autoridade realizadas com o *termo* $t_k \in T_j$. Percorrendo cada elemento da lista R_k verifica-se o valor do *PageRank* à Priori de cada autoridade a_i categorizada com

[6] http://jung.sourceforge.net/

o termo t_k e então adiciona-se esse valor ao *boost* de cada documento categorizado com o termo t_k pela autoridade a_i. O valor de pr_i adicionado ao *boost* de cada documento é representado em ponto flutuante, normalizado com um valor entre 0 e 1.

Algoritmo 5. Indexação dos documentos com o cálculo do fator de *boost*.

Entrada: Lista D com os documentos a serem indexados
Saída: Índice contendo os *boosts* com o valor de *PageRank* das autoridades

Inicializar o *boost* de cada termo em cada documento com o valor 1.0
para cada *documento* $d_j \in D$ **faça**
  $T_j \leftarrow termos(d_j)$
  para cada $t_k \in T_j$ **faça**
    $R_k \leftarrow \{(u_j, t_k, a_i, p_l) \in Z\}$
    para cada $a_i \in R_k$ **faça**
      se *documento* d_j *foi categorizado por* a_i **então**
        $boost_{k,j} \leftarrow boost_{k,j} + pr_i$
        se d_j *está indexado* **então**
          Atualizar o $boost_{k,j}$ no índice
        fim
        senão
          Indexar d_j
          Atualizar o $boost_{k,j}$ no índice
        fim
      fim
    fim
  fim
fim

A abordagem utilizada para indexação dos documentos levou em consideração quais autoridades nos assuntos relacionados a esses documentos os categorizaram. Por exemplo, no momento da indexação, um documento sobre `information` recebeu um *boost* das autoridades nesse assunto (isto é, as autoridades que foram categorizadas com a *tag* `information`). Assim, os documentos disponibilizados por autoridades cognitivas em termos de uma consulta podem ser priorizados no *ranking*.

A utilização de um fator de *boost* em cada dimensão do vetor que representa um documento forneceu uma vantagem também com relação ao fator de normalização do *score* dos documentos. Quando a abordagem de cálculo do *PageRank* das autoridades é feito em tempo de consulta, isto é, quando calcula-se um fator de *boost* dos documentos após a indexação, é necessário ajustar o *score* dos documentos recuperados com alguma abordagem de *scoring* (possivelmente *tf-idf*). Dessa forma, tem-se um cálculo de importância das autoridades (realizado por meio de *PageRank*) cuja escala precisa ser "ajustada" antes

de ser combinada com o cálculo da importância dos recursos. Durante os testes realizados neste trabalho, pode-se perceber que esse "ajuste" de escala pode consumir recursos de tempo para recuperar as informações no índice.

4.4 Discussão

Os conceitos introduzidos pela chamada *Web Social* trouxeram novas perspectivas e oportunidades para a RI, especialmente no sentido de aproveitar o conhecimento gerado pelos usuários dos sistemas *Web* como parte da determinação dos resultados de busca e recomendação (Abel et al., 2008). Pereira e da Silva (2008b) sugere que a utilização de uma abordagem que considera a própria opinião dos usuários de um SRI com relação à autoridade cognitiva das fontes de informação. Neste trabalho, é feita a hipótese de que a consideração das autoridades cognitivas para o cálculo do *score* melhora a recuperação de documentos relevantes e com qualidade. Para avaliar essa hipótese, optou-se por modelar um SRI com base na abordagem de *Folkauthority* e por simular as concessões de autoridade nesse sistema, devido a ausência de uma rede social com as características desejadas.

Em um SRI o contexto[7] em que um usuário se encontra pode influenciar a avaliação da relevância e da qualidade dos documentos recuperados e, portanto, juntamente com o modelo do usuário devem influenciar o *ranking* das informações. No caso da dimensão da autoridade cognitiva, diversos autores que pesquisaram o tema apontam que a autoridade cognitiva é uma questão subjetiva e situacional (Fritch e Cromwell, 2001; Metzger, 2007; Rieh, 2002; Russel, 2005; Wilson, 1983), sugerindo que fatores tais como o conhecimento prévio sobre o assunto por parte de quem categoriza as autoridades e o grau de proximidade entre a autoridade e quem concede autoridade influenciam o processo de concessão de autoridades cognitivas, podendo ser considerados fatores contextuais. Do ponto de vista do sistema, uma *tag* que foi utilizada para descrever um documento em determinado contexto pode não ser apropriada em outro contexto (Sanderson, 2010). Um exemplo típico é o vocabulário utilizado pelo usuário para expressar sua necessidade de informação: nem sempre o vocabulário daqueles usuários que realizaram a categorização dos documentos é semelhante ao vocabulário do usuário que realiza a busca.

Com relação ao peso das autoridades, pode-se afirmar que o cálculo do *PageRank* à Priori (White e Smyth, 2003) resume uma métrica de autoridade dentro da cadeia de

[7] O termo "contexto" neste trabalho refere-se a modelos que caracterizem a situação em que um usuário faz uso da informação como, por exemplo, *quando* e *como* ele a utiliza. Exemplos de informação de contexto incluem o modelo do usuário, a localização do usuário, o histórico de consultas, o horário da busca, o gênero do documento, etc. (Melucci, 2008)

autoridades, levando em consideração a popularidade dessas autoridades (Page et al., 1998). Essa característica está de acordo com aquilo que foi discutido em (Pereira e da Silva, 2008b), sobre a propagação e a retenção de autoridade na cadeia de autoridades. A característica de peso das arestas no algoritmo *PageRank* à Priori foi levada em consideração na Equação 4.1, com a diferença de que constante β, de reinicialização do percurso na rede desde um nó aleatório, não estar demonstrada na equação (apesar da constante existir nos cálculos realizados, pois o grafo da rede considerada pode possuir componentes desconexos).

Com relação ao cálculo do *ranking* dos recursos, na abordagem descrita neste trabalho foi utilizado um fator de *boost* introduzido em cada termo de um documento em tempo de indexação. Conforme comentado, essa abordagem de cálculo em tempo de indexação trouxe vantagens com relação ao tempo de execução de uma consulta e com relação a normalização das escalas do cálculo da importância das autoridades e dos documentos. No entanto, para que essa abordagem seja efetiva em um sistema de larga escala com usuários reais, é necessária uma maneira eficiente de atualizar o *boost* dos termos de índice de cada documento.

Por fim, a simulação da cadeia de autoridades foi realizada tendo em mente os modelos de SBFs apresentados na literatura (Abel et al., 2008; Golder e Huberman, 2006; Hotho et al., 2006). Dessa forma, os passos realizados na simulação estão de acordo com o que os pesquisadores indicam como sendo a dinâmica dos SBFs. Halpin et al. (2007), Golder e Huberman (2006) e Sen et al. (2006) analisam extensamente dados provenientes de diferentes SBFs e discutem as questões de distribuição de *tags* na personomia dos usuários e nos documentos, de interseção de vocabulário de usuários dos sistemas e de quantidade de *tags* utilizadas em cada categorização. Dessa forma, considerando que o arcabouço especificado pelo conceito de *Folkauthority* está ancorado na utilização de folksonomia, espera-se que a simulação realizada sobre a concessão de autoridades possa ter resultado em um modelo instanciado que se assemelhe bastante ao que seria um sistema baseado no conceito de *Folkauthority*.

No próximo capítulo serão apresentados os testes e os resultados dos testes realizados sobre os dados obtidos com o emprego da metodologia descrita neste capítulo. Essa apresentação será seguida de uma discussão sobre os resultados obtidos e de apontamentos de trabalhos futuros, dadas as conclusões retiradas dos resultados.

Capítulo

5

Avaliação da Proposta

Nos capítulos anteriores foram apresentadas a proposta de um SRI que utiliza a abordagem de *Folkauthority*, denominado de *AuthoritySearch*, e o objetivo de verificar se a utilização dessa abordagem para o cálculo do *score* dos documentos apresenta melhorias para a realização da tarefa de RI, quando comparados com um cálculo tradicional. Este objetivo é baseado na hipótese de que é possível indicar as autoridades cognitivas das fontes de informação por meio de *tags*, de forma que as informações provenientes dessas autoridades sejam priorizadas no momento da recuperação. Neste capítulo serão apresentados os resultados com relação à validação dessa hipótese, os quais são discutidos com base em um teste de significância estatística (Smucker et al., 2007) sobre os vetores de NDCG (Jarvelin e Kekalainen, 2002) gerados a partir de avaliação da relevância e da qualidade (Saracevic, 2007) dos documentos recuperados.

Para atingir o objetivo relacionado, optou-se por realizar um estudo utilizando a opinião de usuários cotidianos de SRIs a respeito dos resultados de busca no sistema *AuthoritySearch*. Para tanto, os modelos definidos na simulação (Capítulo 4) foram utilizados para popular o sistema *AuthoritySearch* e um cenário de testes com usuários foi executado. O resultado das avaliações realizadas pelos usuários forneceram um conjunto de julgamentos a respeito da relevância e da qualidade de uma série de documentos presentes no índice, definida com base nas necessidades de informação e nas consultas utilizadas nos cenários de busca. A partir da avaliação desses resultados, foi possível atingir os objetivos do estudo, a dizer: i) verificar se há diferença com relação às dimensões da relevância e da qualidade da informação apresentada por diferentes *rankings*

que consideram autoridades cognitivas e; ii) avaliar quantitativamente a utilização de *Folkauthority* para o *ranking* dos documentos, verificando quanto pode ser melhor (ou pior) essa abordagem em comparação com a abordagem tradicional de *tf-idf* para RI.

Para proceder com a comparação das diferentes abordagens para RI, foi necessário realizar as seguintes tarefas: i) a definição da *coleção de consultas*, a qual foi definida com base na necessidade de informação dos usuários participantes e, ii) o *julgamento dos documentos* com relação aos critérios de qualidade e relevância, a foi realizada por usuários participantes[1]. O resultado da realização dessas tarefas gera indicadores sobre o desempenho das diferentes abordagens para RI, os quais foram utilizados para atingir os objetivos estabelecidos.

5.1 Critérios para avaliação

Ao se realizar a avaliação de um SRI é necessário definir um conjunto de critérios a serem empregados, os quais estão relacionados às características das informações recuperadas. Dentre os possíveis critérios, escolheu-se os conceitos de relevância e de qualidade, os quais foram discutidos por diversos autores (Barry, 1994; Goffman, 1964; Greisdorf, 2000; Kagolovsky e Mohr, 2001; Kargar, 2011; Knight e Burn, 2005; Lachica et al., 2008; Saracevic, 1975; Saracevic et al., 1988) em estudos sobre os critérios utilizados por usuários de SRIs para o julgamento de informações, conforme foi descrito no Capítulo 2, Seção 2.3. Procurando uma definição clara para os conceitos de relevância e qualidade observa-se que a relevância, no contexto da RI, é também conceituada como a medida com que a informação é transmitida por um documento dada uma consulta (Goffman, 1964). Isto é, a relevância é avaliada de acordo com uma necessidade de informação, de forma que um documento é relevante se contém informação que satisfaça esta necessidade de informação. Lachica et al. (2008) relacionou o conceito de *relevância* à satisfação da necessidade de informação de um usuário, enquanto a *qualidade* da informação foi relacionada com o valor intrínseco que essa informação traz ao usuário. Barry (1994) identificou os seguintes critérios utilizados pelos participantes para julgarem as informações com as quais interagiram: extensão do conhecimento do usuário sobre a fonte e/ou o conteúdo da informação, habilidade para compreensão do documento e validade subjetiva (extensão na qual o usuário concorda com a informação).

Com base nas definições estudadas e, tendo em vista a necessidade da definição desses conceitos de forma a serem compreensíveis por usuários habituais de SRIs, os

[1]Os avaliadores dos resultados de busca, os quais participaram da pesquisa verificando os resultados e respondendo os formulários de avalição, serão chamados de *usuários participantes* neste trabalho.

quais não necessariamente possuam conhecimento técnico sobre RI, as seguintes noções de "qualidade" e "relevância" foram passadas aos usuários participantes da pesquisa:

- **Relevância** é a medida na qual a informação contida em um documento é útil para a resolução do(s) problema(s) elencado(s) no cenário de busca proposto.
- **Qualidade** é a medida na qual o participante julga que a informação contida em um documento seja bem apresentada, abrangente, profunda e completa, sendo considerado um contexto mais amplo do que o cenário de busca para a realização de tal julgamento.

A fim de quantificar tais critérios, são empregadas medidas como a precisão (Manning et al., 2009), as quais são realizadas a partir de julgamentos binários exclusivos sobre a relevância e a qualidade das informações (relevante ou não-relevante). No entanto, vários autores concordam que a noção de relevância é formada por diversas dimensões, dentre as quais inclui-se a da *medida* (ou grau) de relevância, representada em termos de um julgamento comparativo com um julgamento absoluto sobre a relevância/qualidade de um item de informação. Uma vez que os critérios de qualidade e de relevância se expressam melhor com uma métrica relativa (isto é, gradual), a noção de "medida" (ou extensão) da relevância/qualidade foi trazida a este trabalho, permitindo ao usuário participante aferir um grau de relevância/qualidade a cada documento no formulário de avaliação dos resultados de busca.

Outro fato a observar é que devido aos dados utilizados no estudo serem incompletos (isto é, somente uma parte dos recursos presentes na base de dados foram avaliados), é impossível o cálculo da métrica de *cobertura* (ver Equação 2.7) para as informações recuperadas no sistema *AuthoritySearch*, uma vez que essa métrica é baseada na quantidade total de documentos presentes no índice que são relevantes à uma consulta. A avaliação de todos os recursos presentes no índice é uma tarefa custosa (Bailey et al., 2003; Sanderson, 2010), sendo inviável efetivá-la no período de realização deste trabalho. Sendo assim, foi necessário utilizar uma métrica para a avaliação de SRIs que fosse baseada na avaliação da relevância dos n primeiros resultados de uma busca. Com o termo "primeiros" deseja-se também ressaltar a noção de ordem dos resultados de busca, comparados de acordo com sua importância para uma determinada consulta.

Os critérios escolhidos neste trabalho e as restrições impostas pelo conjunto de dados levam a necessidade de se escolher métricas que transformem a avaliação gradual da relevância e da qualidade de documentos em números. Para tal, foi escolhida a métrica chamada de NDCG (Jarvelin e Kekalainen, 2002), a qual é baseada no CG (*Cumulated*

Gain - Ganho Cumulativo) dos n primeiros resultados de uma busca em um tópico e foi apresentada no Capítulo 2. O CG é a soma do ganho (isto é, do valor aferido na escala de relevância utilizada para julgamento das informações) de cada documento do *ranking*, desde o documento na posição 1 até o documento na posição n. Essas métricas se mostram adequadas pelo objetivo do estudo ser baseado na avaliação de um algoritmo de *ranking* das informações e também por ser baseado na utilização de uma noção de *graus de relevância* para a avaliação dos documentos presentes no conjunto de dados de teste. A métrica de NDCG está relacionada com um nível ideal de precisão dos resultados de uma busca (Jarvelin e Kekalainen, 2002; Manning et al., 2009), o que permite a comparação da magnitude da diferença entre duas curvas DCG para diferentes abordagens de *ranking* e a aplicação de testes de significância estatística sobre os dados gerados com a métrica.

5.2 Coleta de dados

Neste trabalho, os usuários avaliaram os resultados de busca a partir de diferentes necessidades de informação expressas por diferentes consultas passadas ao sistema. No Capítulo 2, Seção 2.3 foram discutidas algumas questões sobre a definição das consultas a serem realizadas pelos usuários participantes e foram também descritas duas maneiras de se obter tais consultas, uma de forma aleatória (com base em um *log* de consultas, por exemplo) e a outra utilizando a necessidade de informação do usuário participante. Com base nessas questões, foi especificado que na metodologia adotada nesta pesquisa os usuários participantes seriam responsáveis por exprimir suas necessidades de informação aos pesquisadores e por guiar a elaboração das consultas a serem realizadas ao sistema. Deste modo, cada consulta foi definida com base em um cenário (que descreve a necessidade de informação), preparado após uma conversa informal com o usuário participante na qual esse expunha alguns de seus principais assuntos de interesse. Os cenários foram elaborados de forma a permitir a expressão de consultas por meio de termos frequentemente presentes nos dados gerados na simulação (isto é, nos termos presentes no índice do sistema), os quais são, em sua maioria, termos relacionados à tecnologia e à *Web*.

Para cada consulta realizada pelos usuários participantes, foi solicitado que se realizasse a recuperação de informação utilizando dois mecanismos de busca, aqui denominados de busca "Verde" e busca "Amarela", as quais utilizam esquemas de *ranking* distintos e, consequentemente, exibem resultados de busca ordenados de forma diferente. A busca "Verde" utiliza o cálculo de *score* dos documentos considerando o *boost* introduzido pelas autoridades, enquanto a busca "Amarela" utiliza um esquema de *scoring* por *tf-idf* apenas, no entanto nenhum dos usuários tinha conhecimento desse fato. A Figura 5.1 ilustra essas

etapas. A linha tracejada indica que as etapas relacionadas devem ser repetidas tanto para a busca “Verde” quanto para a busca “Amarela”.

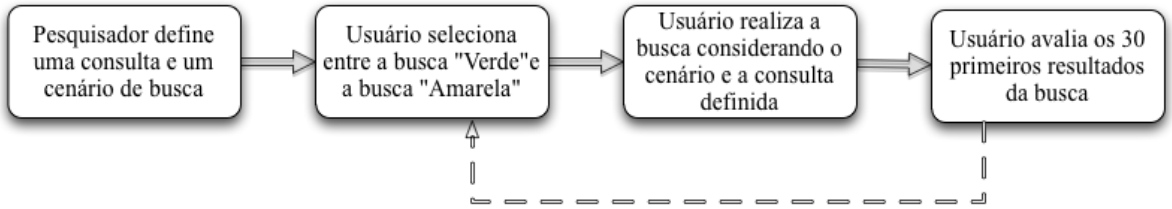

Figura 5.1: Etapas dos testes realizados.

Nos estudos realizados nesta pesquisa, foram utilizados 7 usuários participantes, os quais realizaram algumas consultas e avaliaram os 30 primeiros resultados de busca para cada uma delas. Os participantes foram selecionados de acordo com dois critérios: a necessidade de que o usuário utilizasse SRIs cotidianamente (todos os participantes utilizam pelo menos mecanismos de busca na *Web*) e a necessidade de que o usuário possuísse fluência de leitura na língua inglesa. Foram utilizados 9 tópicos diferentes, distribuídos entre os usuários. Esses tópicos estão demonstrados na Tabela 5.1, a qual relaciona cada tópico com o respectivo usuário que realizou a avaliação. Cada tópico é representado por uma necessidade de informação e sua respectiva consulta.

Cada uma das consultas descritas na Tabela 5.1 foi realizada no sistema *Authority-Search* utilizando a busca “Verde” e a busca “Amarela”, de forma a gerar uma avaliação para os resultados dos três esquemas de *ranking* descritos no Capítulo 4, Seção 4.3.1. Na busca “Verde” são consideradas as autoridades cognitivas para o *ranking* dos documentos, de forma que o *score* de cada documento d_j dada uma consulta $\vec{q}$ (denominado de $s_{j,q}$) é calculado com base na Equação 4.1, a qual possui distintamente duas componentes: uma relacionada ao cálculo de $tf_{i,j} \times idf_i$ e uma relacionada ao cálculo de $boost_{i,j}$, considerando um termo t_i em um documento d_j. Quando um usuário participante avalia um documento d_j utilizando a busca “Verde” são armazenados (além das informações acerca da avaliação) os valores de $s_{j,q}$ e da soma $\sum_{t_i \in q}(tf_{i,j} \times idf_i)$ para esse documento. Dessa forma, tem-se duas maneiras distintas de ordenar os documentos produzidos nessa consulta. A primeira utiliza o esquema *AuthorityRank* e considera somente $s_{j,q}$, realizando a ordenação dos documentos de acordo com esse valor. A segunda utiliza o esquema AR/*tf-idf* (“*tf-idf* com filtro de autoridades”), o qual considera uma reordenação dos 30 primeiros documentos recuperados utilizando o esquema *AuthorityRank*, reordenando os resultados com base na componente $\sum_{t_i \in q}(tf_{i,j} \times idf_i)$ do valor de $s_{j,q}$. Assim sendo, temos dois *rankings*

Necessidade de informação	**Consulta**	**Usuário participante**
O usuário deseja desenvolver um sistema para a Web 2.0 e procura informação sobre bibliotecas e API's que auxilie na utilização de ferramentas AJAX.	ajax api	Usuário 1
O usuário necessita realizar um teste de usabilidade em um software e procura informação sobre procedimentos, técnicas, dicas e tutoriais para a realização de um teste de usabilidade.	usability test tutorial	Usuário 1
O usuário deseja construir um blog sobre gastronomia e está procurando informação a respeito de tutoriais para a construção de um blog.	how to tutorial build make blogs	Usuário 2
O usuário está desenvolvendo a interface de um sistema e deseja informação sobre especificações e exemplos de utilzação de padrões de projeto de interfaces em design de interação.	interaction design patterns user interface UI	Usuário 3
O usuário necessita infomação sobre modelagem e desenvolvimento com Orientação a Objetos (OO) utilizando ferramentas da linguagem UML.	object oriented oo oop uml unified modeling language tool	Usuário 3
O usuário deseja informação sobre metodologias para a engenharia de ontologia.	ontology engineering methodology	Usuário 4
O usuário necessita realizar uma revisão sobre abordagens/técnicas para recomendação de informação em sistemas baseados em folksonomia.	tagging recommendation personomy folksonomy	Usuário 5
O usuário busca informação sobre a descrição e exemplos de utilização dos elementos da linguagem HTML5.	html5 elements description	Usuário 6
O usuário busca conceitos, definições e exemplos de pesquisas na área de visualização e recuperação de informação.	information retrieval visualization	Usuário 7

Tabela 5.1: Relação entre os tópicos avaliados e os usuários participantes

distintos, ambos considerando as autoridades cognitivas para mostrar seus resultados de

busca, sendo que um deles é simplesmente uma reordenação dos resultados com base na frequência dos termos da consulta nos documentos.

A busca "Amarela" utiliza o esquema tradicional chamado de *tf-idf* para realizar a ordenação dos documentos. Assim, pode-se avaliar duas abordagens diferentes para considerar autoridades cognitivas no *ranking* bem como avaliar abordagens que utilizam e que não utilizam *Folkauthority* para o *ranking* dos recursos, de acordo com os objetivos que foram determinados no estudo descrito.

Uma questão relevante sobre o estudo realizado diz respeito a quantidade de pontos na escala de relevância e da qualidade dos documentos, os quais podem ser aferidos pelos usuários participantes. Alguns autores discutem sobre a quantidade de pontos que deve conter essa escala. No trabalho de Yao (1995) é realizada uma revisão sobre esse assunto. O autor cita os trabalhos de Keen (1971) e de Saracevic et al. (1988), que utilizam em seus estudos escalas de quatro e de três pontos, respectivamente. Vakkari e Sormunen (2004) também realizaram um trabalho no qual verificou-se o efeito da gradação da relevância nos resultados de avaliação de um SRI interativo. No entanto, conforme destaca Yao (1995), há uma carência de pesquisas com relação a esse tema.

Neste trabalho foi utilizado uma escala de relevância e de qualidade dos documentos contendo cinco pontos, variando do grau 0 ao grau 4. Além disso, o documento entregue aos usuários participantes com instruções para a realização das avaliações contém uma *sugestão* de que os graus da escala sejam aferidos de acordo com as premissas estabelecidas na Tabela 5.2. Nas linhas referentes à coluna "Premissas" estão descritas as premissas para a relevância e para a qualidade dos documentos. Nessa tabela, verifica-se também que foi sugerido ao usuário participante que um documento fosse avaliado com grau 0 quando fosse totalmente irrelevante ou não tivesse nenhum indicativo de qualidade. Foi sugerido ao usuário que considerasse indicativos de qualidade como sendo características tais como boa apresentação, veracidade, abrangência e/ou profundidade com relação ao assunto e a utilidade da informação para o usuário. Um documento deveria ser aferido com o grau 4 caso fosse altamente relevante, satisfazendo todos os tópicos relacionados na necessidade de informação ou fosse de excelente qualidade, possuindo todos os aspectos de qualidade indicados. Os graus intermediários (1, 2 e 3) estão descritos na tabela.

Os recursos do sistema *AuthoritySearch* incluem uma interface para a busca típica de mecanismos de busca da *Web*, na qual é possível selecionar o esquema de busca (busca "Verde" ou busca "Amarela"). Além disso, cada resultado de busca trás uma opção de interface para a avaliação dos documentos recuperados, na qual é possível aferir o grau de relevância e de qualidade desses documentos. No documento elaborado como roteiro para a avaliação dos resultados de busca (vide Apêndice D) foi sugerido ao usuário participante

Grau da escala	Premissa
0	Documento irrelevante, não contém nenhuma informação sobre o tópico relacionado *e* Documento de péssima qualidade, não possui nenhum indicativo dos aspectos de qualidade.
1	Pouco relevante, o documento apenas aponta para o tópico. Não satisfaz nenhum tópico relacionado à necessidade de informação *e* Baixa qualidade, o documento possui uma apresentação indesejável, conteúdo pouco abrangente ou possui fontes duvidosas.
2	Marginalmente relevante, o documento contém alguma informação sobre o tópico, no entanto não se esgota em nenhuma das informações. Satisfaz apenas parte dos tópicos relacionados à necessidade de informação *e* Média qualidade, o documento possui alguns indicativos de qualidade, tais como uma apresentação desejável ou uma fonte confiável, no entanto não preenche mais do que dois indicativos de qualidade.
3	Relevante, satisfaz todos os tópicos relacionados à necessidade de informação no entanto não é exaustivo nas informações sobre esses tópicos *e* Boa qualidade, o documento possui mais de dois indicativos de qualidade, possuindo informação clara e abrangente sobre o assunto de que trata.
4	Altamente relevante, satisfaz todos os tópicos relacionados à necessidade de informação e apresenta conceitos e exemplos sobre esses tópicos *e* Excelente qualidade, o documento possui todos os indicativos de qualidade.

Tabela 5.2: Descrição das sugestões para consideração dos graus de relevância e de qualidade

que considerasse as descrições e as escalas para os critérios que estão demonstradas na Tabela 5.3.

Para realizar a tarefa de avaliação, cada usuário utilizou o formulário fornecido no sistema *AuthoritySearch*, o qual é possível observar na Figura 5.2. Nesse formulário, o usuário possui a opção de marcar qual o grau de relevância e de qualidade ele considera que o documento possua.

Por fim, cabe comentar algumas precauções tomadas neste estudo sobre os efeitos no julgamento das informações por parte dos usuários participantes. Saracevic et al. (1988) afirmam que o contexto influencia muito na tarefa de RI e de julgamento das informações.

Descrição do critério	Escala
Em que extensão o conteúdo apresentado pelo recurso é relevante ao seu interesse de busca ?	Nenhuma (0), Pouca (1), Média (2), Bastante (3), Total (4)
Como você avalia a qualidade (quanto a completude, profundidade e abrangência) do conteúdo apresentado pelo recurso ?	Péssima (0), Ruim (1), Mediana (2), Boa (3), Excelente (4)

Tabela 5.3: Descrição dos critérios para avaliação das informações e suas respectivas escalas

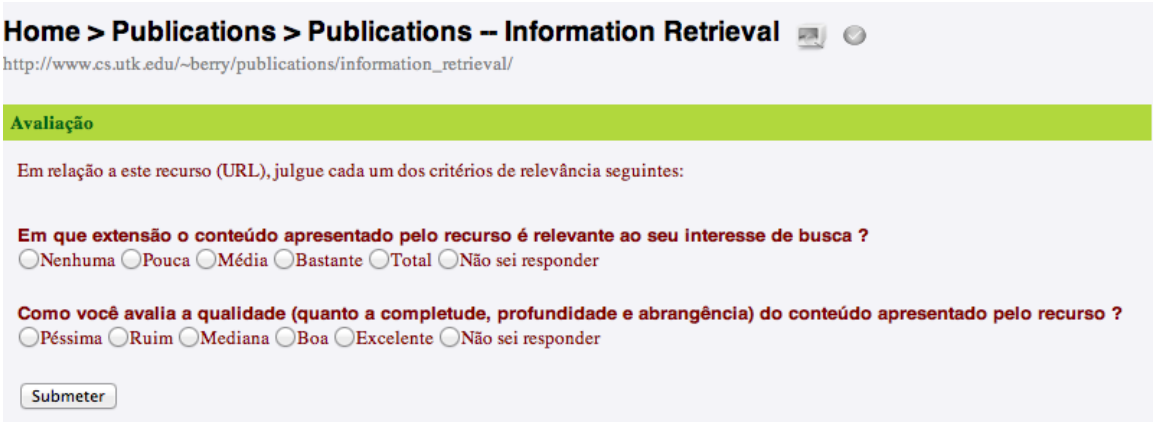

Figura 5.2: Formulário para avaliação dos resultados de busca.

Contexto, segundo o autor, pode ser descrito por um conjunto de variáveis que afetam os eventos relacionados à tarefa de RI e pode ser originário de aspectos ambientais (externos) ou cognitivos (internos). Nesse trabalho, o autor examina quatro aspectos relacionados ao contexto chamados de *problema subjacente*, *intenção*, *estado do conhecimento* e *estimativa de conhecimento público*. O *problema subjacente* diz respeito a um desconhecimento sobre uma situação ou sobre uma tarefa por parte do usuário e significa uma dificuldade em se encontrar uma solução ou uma resposta. A *intenção* diz respeito ao propósito que o usuário possui com relação à informação requisitada/avaliada. As questões que envolvem a intenção dizem respeito, por exemplo, às características desejáveis da informação com relação ao problema subjacente (completude, precisão, confiabilidade, recência, língua, fonte, etc.) e o tempo e o esforço que o usuário está disposto a gastar para absorver ou compreender a informação. O *estado do conhecimento* diz respeito ao grau de conhecimento que o usuário possui a respeito do problema subjacente. Por fim, a *estimativa de conhecimento público* define a expectativa do usuário sobre o conhecimento público, isto é, sobre a informação que está disponível sobre o problema subjacente. Essa questão envolve aspectos como a percepção sobre a organização da informação e do que

se espera obter com as informações. Em função deste aspecto, Saracevic et al. (1988), a fim de tornar o estudo realizado o mais parecido com o "ambiente natural" possível (nas próprias palavras dos autores), determinou que os usuários participantes do estudo realizassem buscas relacionadas à sua pesquisa/trabalho. Essa medida também foi tomada no estudo realizado neste trabalho, uma vez que os tópicos foram definidos com base em uma conversa prévia sobre os assuntos de interesse do usuário participante.

Harter (1996) afirma que a literatura envolvendo estudos experimentais sobre os fatores que influenciam o julgamento da relevância é ampla. O autor cita o trabalho de Schamber (1994), que enumera 80 fatores que influenciam o julgamento da relevância de documentos por parte de seus usuários. Dentre os fatores selecionados por Harter (1996), há aqueles que se relacionam com o estudo deste trabalho, dos quais é possível citar a educação formal e o conhecimento/experiência com relação ao assunto da consulta. Sobre a relação desse fator com o estudo apresentado, pode-se afirmar que todos os participantes tinham educação formal e que nenhum conhecia sobre o assunto de forma aprofundada. Com relação à recência e ao estilo dos documentos, esses foram obtidos a partir da *Web* (todos são documentos representados por uma URL, tais como vídeos, fotos, sítios e documentos de texto) em meados do ano de 2010 até o início de 2011[2]. Com relação ao SRI implementado, pode-se elencar os critérios de esforço e navegabilidade. Essas duas características foram reforçadas no projeto da interface do sistema *AuthoritySearch*, no qual adicionou-se a possibilidade de visualizar rapidamente o conteúdo de um documento utilizando a própria interface de busca. Caso o usuário participante desejasse explorar com maiores detalhes o conteúdo do documento, a interface oferecia a possibilidade de abrir o documento em uma janela separada.

Ainda, com relação às condições impostas ao usuário participante, Harter (1996) destaca os critérios de ordem de apresentação, tempo para julgar a relevância das informações, tamanho do conjunto de documentos e definição do conceito de relevância por parte do usuário. Com relação à ordem de apresentação, todos os documentos foram apresentados de acordo com o *ranking* da abordagem de RI utilizada, exceto a abordagem AR/*tf-idf*, que é somente uma reordenação do esquema de *ranking* AR apresentado ao usuário. Sobre o tempo dado ao usuário para julgar a relevância das informações, todos os usuários tiveram tempo à disposição para a realização dos testes. O tamanho do conjunto de documentos a serem avaliados foi limitado a 30 por consulta, de forma que não fosse tão custoso ao usuário categorizar duas consultas, além de representar as três primeiras páginas de resultados de um mecanismo típico de busca na *Web*. No entanto,

[2]Os dados foram obtidos a partir de experimentos que executaram entre o período de Março de 2010 a Março de 2011.

neste trabalho não houve uniformidade com relação ao número de documentos avaliados por usuário, já que cada um realizou um número diferente de consultas. Ainda, a definição dos conceitos de relevância e qualidade foi previamente explicada a cada usuário. Além disso, ao usuário era permitida a consulta ao documento que explicava o procedimento de avaliação, o qual continha a definição utilizada para os conceitos de qualidade e relevância da informação.

5.3 Análise dos resultados

Após a coleta dos dados das avaliações realizadas pelos usuários, foi possível comparar os três esquemas de *ranking* utilizados neste trabalho, a dizer: *AuthorityRank*, *tf-idf* e AR/*tf-idf*. Para cada consulta q_m realizada, foi gerado um conjunto chamado de O_{q_m}, constituído de seis vetores com valores de NDCG-10. Optou-se pelo NDCG-10 pois ele representa a persistência de um usuário analizar 10 resultados de busca, que é o tamanho padrão da página de resultados de um mecanismo de busca na *Web* atual. Para cada esquema de *ranking* r_n, produziu-se dois vetores $\vec{o}_{q_m,c_{rel},r_n}$ e $\vec{o}_{q_m,c_{qual},r_n}$ de 30 dimensões cada, contendo respectivamente os valores de NDCG-10 sobre o critério de relevância (c_{rel}) e sobre o critério de qualidade (c_{qual}) dos 30 primeiros documentos para a consulta q_m e o esquema de *ranking* r_n. A busca "Verde" gerou os vetores para os esquemas de *ranking AuthorityRank* ($\vec{o}_{q_m,c_{rel},AR}$ e $\vec{o}_{q_m,c_{qual},AR}$) e AR/*tf-idf* ($\vec{o}_{q_m,c_{rel},AR/tf\text{-}idf}$ e $\vec{o}_{q_m,c_{qual},AR/tf\text{-}idf}$), enquanto a busca "Amarela" gerou os vetores para o esquema *tf-idf* ($\vec{o}_{q_m,c_{rel},tf\text{-}idf}$ e $\vec{o}_{q_m,c_{qual},tf\text{-}idf}$), totalizando seis vetores por consulta. Definiu-se também o conjunto O, contendo todos os elementos dos conjuntos O_{q_m} para cada consulta q_m utilizada.

A Tabela 5.4 mostra um exemplo hipotético de vetores do conjunto $O_{\texttt{text}}$ com cinco posições para valores de NDCG-2. O exemplo ilustrado na tabela representa uma consulta contendo o termo `text`, na qual foram avaliados os 5 primeiros resultados para os dois esquemas de busca, gerando os seis vetores do conjunto $O_{\texttt{text}}$. Além disso, a Figura 5.3 mostra graficamente a comparação entre os vetores de NDCG-2 referentes ao critério de relevância para os três esquemas de *ranking* definidos na Tabela 5.4.

No Apêndice A estão anexos os resultados em termos de NDCG-10 para os critérios de relevância e de qualidade dos nove tópicos avaliados pelos usuários participantes, os quais são apresentados conforme o modelo considerado na Figura 5.3. As médias para os valores de NDCG-10 dos nove tópicos avaliados são apresentadas nas Figuras 5.4 e 5.5, as quais consideram, respectivamente, os critérios de relevância e de qualidade dos documentos.

Consulta (q_m)	**Critério** c_{rel}			**Critério** c_{qual}		
	AR	AR/*tf-idf*	*tf-idf*	AR	AR/*tf-idf*	*tf-idf*
`text`	1,000	0,667	0,500	1,000	0,333	0,250
	0,500	0,333	0,250	0,500	0,167	0,143
	0,556	0,444	0,182	0,625	0,375	0,100
	0,455	0,455	0,154	0,500	0,400	0,091
	0,615	0,462	0,154	0,583	0,417	0,091

Tabela 5.4: Modelo para descrição dos resultados com base em vetores de NDCG para uma consulta

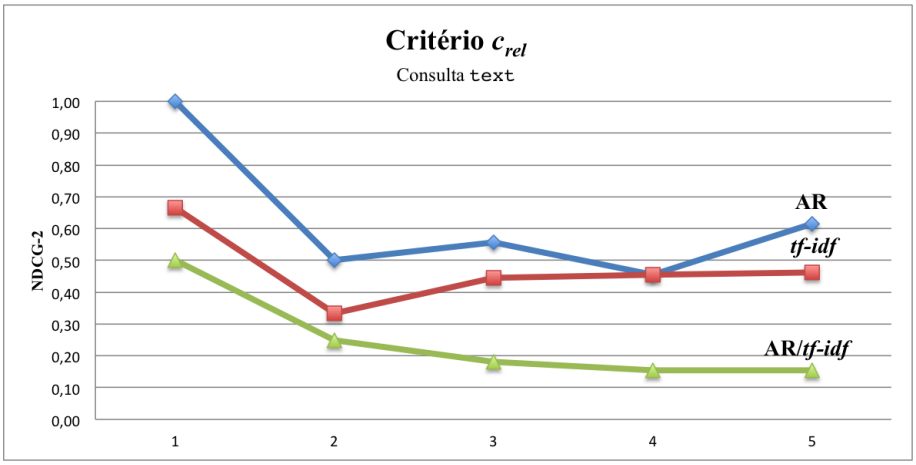

Figura 5.3: Gráfico para os valores de NDCG mostrados na Tabela 5.4.

As médias de NDCG-10 para cada um dos esquemas de *ranking* separadas por *faixas de resultados* são mostradas nas Tabelas 5.5 e 5.6, nas quais são destacados em negrito os maiores valores de NDCG-10 para as faixas dos 5, 10, 20 e 30 primeiros resultados (o Apêndice B trás anexo a média de NDCG-10 para cada tópico, divido por faixas de resultados e por esquema de *ranking*). Observando somente as médias apresentadas, é possível afirmar que o esquema de *ranking* *AuthorityRank* apresenta resultados melhores para todas as faixas de resultados de busca. Porém, para compreendermos melhor estes resultados, é necessário uma análise estatística dos mesmo. Para tanto, foi aplicado um teste *T de Student* unicaudal entre os pares de vetores gerados pelos esquemas de *ranking*, procurando determinar se houve uma melhoria significativa entre eles.

As Tabelas 5.7 e 5.8 mostram as faixas de resultados para os critérios de relevância e de qualidade sobre as quais pode-se, ou não, considerar que os resultados apresentados pelos esquemas de *ranking* baseado nas autoridades cognitivas sejam significativamente melhores para um nível de significância de 95% ($\alpha = 0,05$). Esses resultados são apresentados

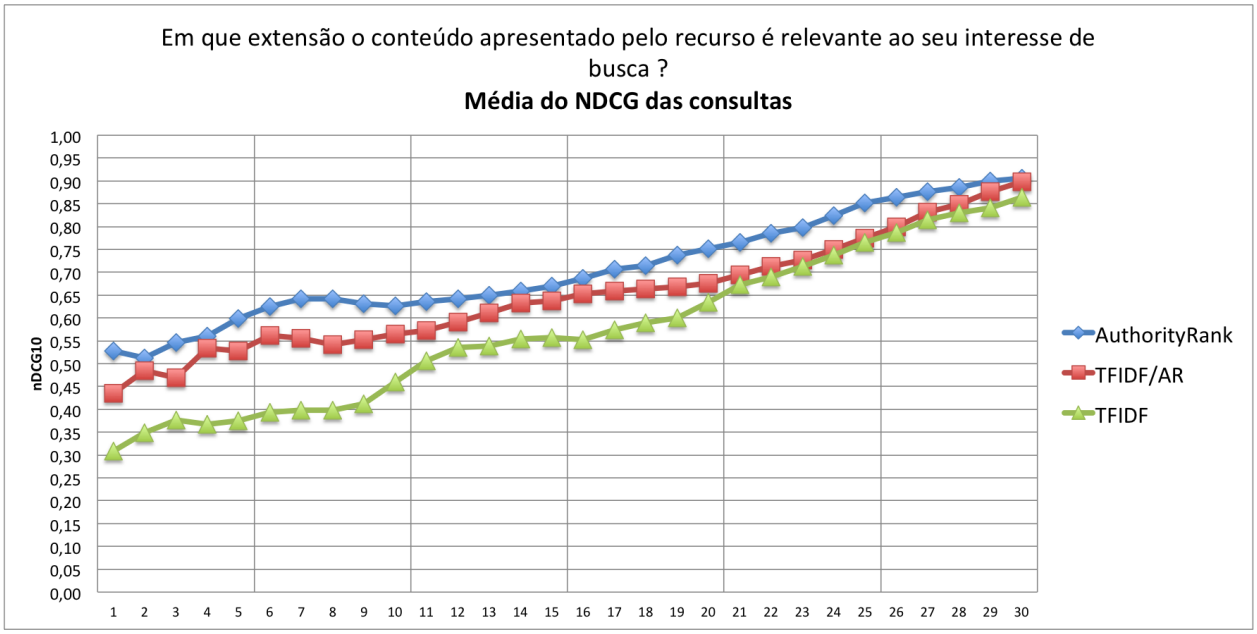

Figura 5.4: Valores da média da métrica NDCG para as consultas realizadas, considerando o critério de relevância dos documentos.

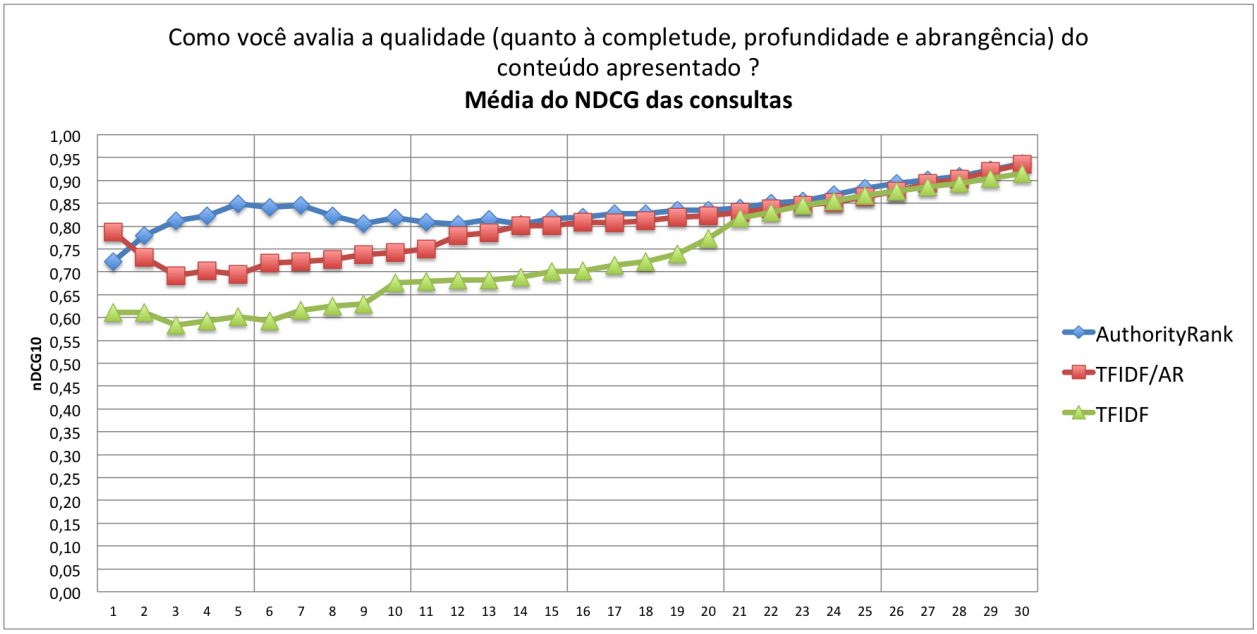

Figura 5.5: Valores da média da métrica NDCG para as consultas realizadas, considerando o critério de qualidade dos documentos.

numericamente no Apêndice C. As tabelas relacionam as faixas de resultados com os pares de vetores de NDCG-10, mostrando se a diferença entre os esquemas é ou não significativa.

Rank	**Média**			**Desvio padrão**		
	AR	AR/TF-IDF	TF-IDF	AR	AR/TF-IDF	TF-IDF
Cinco primeiros	**0,549**	0,490	0,355	0,033	0,041	0,029
Dez primeiros	**0,591**	0,523	0,384	0,050	0,044	0,040
Vinte primeiros	**0,638**	0,580	0,474	0,066	0,070	0,100
Trinta primeiros	**0,707**	0,650	0,573	0,116	0,123	0,168

Tabela 5.5: Média dos valores de NDCG, considerando o critério de relevância dos documentos

Rank	**Média**			**Desvio padrão**		
	AR	AR/TF-IDF	TF-IDF	AR	AR/TF-IDF	TF-IDF
Cinco primeiros	**0,797**	0,722	0,600	0,049	0,040	0,013
Dez primeiros	**0,812**	0,726	0,614	0,038	0,028	0,026
Vinte primeiros	**0,816**	0,762	0,661	0,027	0,045	0,056
Trinta primeiros	**0,839**	0,800	0,730	0,044	0,068	0,111

Tabela 5.6: Média dos valores de NDCG, considerando o critério de qualidade dos documentos

Análisando a Tabela 5.7, a qual refere-se ao critério de relevância, pode-se verificar que: i) o esquema AR é superior ao esquema *tf-idf* em todas as faixas de resultado, ii) não existe diferença significativa entre os esquemas AR e AR/*tf-idf* acima dos 10 resultados e, mesmo nos 10 resultados, a diferença é pequena e, iii) o esquema AR/*tf-idf* é melhor que o esquema *tf-idf* somente até o vigésimo resultado.

	Significância nas faixas de resultados – Critério de relevância					
	AR	*AR/tf-idf*	*AR*	*tf-idf*	*AR/tf-idf*	*tf-idf*
Cinco primeiros	Significativo		Significativo		Significativo	
Dez primeiros	Significativo		Significativo		Significativo	
Vinte primeiros	Não significativo		Significativo		Significativo	
Trinta primeiros	Não significativo		Significativo		Não significativo	

Tabela 5.7: Resultado do teste de significância considerando o critério de relevância

A Tabela 5.8, a qual refere-se ao critério de qualidade, mostra resultados bastante semelhantes àqueles observados na Tabela 5.8, a qual refere-se ao critério de relevância, nela podemos observar que: i) o esquema AR também é superior ao esquema *tf-idf*, ii) também não existe diferença significativa entre os esquemas AR e AR/*tf-idf* acima dos 10 resultados e, iii) o esquema AR/*tf-idf* é melhor que o esquema *tf-idf* somente até o vigésimo resultado.

	Significância nas faixas de resultados – Critério de qualidade					
	AR	*AR/tf-idf*	*AR*	*tf-idf*	*AR/tf-idf*	*tf-idf*
Cinco primeiros	Significativo		Significativo		Significativo	
Dez primeiros	Significativo		Significativo		Significativo	
Vinte primeiros	Não Significativo		Significativo		Significativo	
Trinta primeiros	Não significativo		Significativo		Não significativo	

Tabela 5.8: Resultado do teste de significância considerando o critério de qualidade

Assim, com relação ao objetivo de avaliar se a utilização do esquema de *ranking* AR pode ser melhor (ou pior) do que a abordagem tradicional de *tf-idf*, pode-se afirmar com um grau de confiança de 95%, que o esquema AR é melhor do que o *tf-idf* nos critérios de relevância e qualidade levando em conta o experimento realizado. Não obstante, o cálculo do *score* dos documentos por meio do esquema de *ranking* AR também utiliza resultados do cálculo do *score* por meio de *tf-idf*. A principal diferença no cálculo dos *scores* por meio dessas duas abordagens reside no acréscimo do *boost* proveniente das autoridades cognitivas que categorizaram os documentos recuperados. Dessa forma, o estudo realizado de fato avaliou a diferença entre utilizar o esquema *tf-idf* na sua forma original e utilizá-lo com o *boost* das autoridades. O cálculo desse *boost* em um SRI baseado em Folksonomia é possível utilizando o arcabouço sugerido pelo conceito de *Folkauthority*. Uma vez que a introdução desse *boost* no cálculo do *rank* dos documentos melhorou a relevância e a qualidade dos resultados de busca, é possível também afirmar que a utilização do conceito de *Folkauthority* tem implicações benéficas para a RI em um SBF.

Com relação ao objetivo de verificar se há diferença entre os critérios de relevância e de qualidade dos documentos recuperados por *rankings* que consideram e que não consideram as autoridades cognitivas, observa-se que na média os esquemas de *ranking* AR e AR/*tf-idf* (que consideram as autoridades cognitivas) apresentaram resultados melhores do que para o esquema de *ranking tf-idf* (que não considera as autoridades cognitivas). No entanto, existem casos isolados para alguns tópicos nos quais o esquema que não considera autoridades cognitivas apresenta resultados melhores (ver Apêndices A e B).

Além disso, ao se avaliar o valor da média para os resultados considerando os critérios de relevância e qualidade, observa-se que o valor de NDCG-10 para o critério de qualidade é consideravelmente maior. Por exemplo, nos vinte primeiros resultados de busca o critério de relevância apresenta os valores da média de NDCG-10 iguais à 0,638, 0,580 e 0,474 para os *rankings* AR, AR/*tf-idf* e *tf-idf*, respectivamente, e o critério de qualidade apresenta os valores iguais à 0,816, 0,762 e 0,661. Para os três esquemas de *ranking*, a diferença entre

os valores da média de NDCG-10 para os dois critérios é uniforme, repetindo-se também com relação ao primeiro, aos cinco, aos dez e aos trinta primeiros resultados de busca.

Outro objetivo do estudo era avaliar a efetividade de diferentes abordagens que utilizem o conceito de *Folkauthority* para a RI. Nesse caso, foram comparados dois *ranks* gerados pelos esquemas AR e AR/*tf-idf*. Ao analisar os resultados do estudo realizado, principalmente os gráficos das Figuras 5.4 e 5.5, é possível perceber duas situações diferentes para os critérios utilizados. No caso do critério de relevância, a diferença entre os valores das médias de NDCG-10 para os esquemas de *ranking* AR e AR/*tf-idf* foi pequena em todas as faixas de resultado de busca, enquanto que para o critério de qualidade essa diferença foi maior (com o AR apresentando resultados significativamente melhores somente nos 10 primeiros resultados de busca).

Com base no que foi verificado e avaliado no trabalho, as observações relacionadas à melhoria dos resultados de busca com diferentes esquemas de *ranking* não se aplicam aos esquemas de *ranking* AR e AR/*tf-idf*. Essas duas abordagens mostraram resultados bastante semelhantes, sendo que para o critério de relevância não foi possível afirmar se os resultados apresentados são originados pelos vieses inerentes ao tipo de estudo realizado ou se realmente os dois esquemas de *ranking* apresentam resultados diferentes. Dessa forma, devido ao custo adicional de reordenação dos resultados trazido no algoritmo para o cálculo do *ranking* AR/*tf-idf* (e também por essa abordagem ter apresentado resultados piores do que para a abordagem *AuthorityRank*) julga-se que a utilização da abordagem AR/*tf-idf* em detrimento à abordagem AR não seja recomendada.

Capítulo

6

Conclusões

A Recuperação de Informação no ambiente da *Web* 2.0 apresenta grandes desafios e oportunidades (Maniu et al., 2011; Vlahovic, 2011; Yong, 2011) as quais estão relacionadas à facilidade de publicação e à consequente quantidade de informação disponível na *Web*. Os sistemas baseados em Folksonomia vêm sendo utilizados na tentativa de auxiliar usuários a organizar seus conteúdos por meio de *tags* (Trant, 2009), com o objetivo de reduzir os efeitos da sobrecarga de informação. No entanto, essa técnica sozinha não resolve os desafios discutidos, especialmente no tocante à qualidade das informações disponíveis. Nos SBFs, o conhecimento do usuário que realiza a categorização dos documentos influencia tanto na elaboração dos esquemas de categorização quanto na qualidade dos documentos que são disponibilizados por esses usuários. Conforme discutido por (Wilson, 1983) e demonstrado no Capítulo 3, a questão da *autoridade cognitiva* tem claras relações com o controle da qualidade das informações disponibilizadas. Nesse contexto, a abordagem denominada de *Folkauthority* (Pereira e da Silva, 2008b) foi definida como um arcabouço para a concessão de autoridades cognitivas o qual estabelece uma forma de se indicar o nível de conhecimento de cada usuário do SRI em determinado assunto por meio de *tags*, sendo essa abordagem diretamente relacionada com a qualidade das informações fornecidas pelos usuários.

A hipótese estabelecida neste trabalho foi a de que a indicação do conhecimento de cada fonte de informação pode ser utilizada para priorizar informações no momento da recuperação, melhorando, assim, a qualidade das informações recuperadas em uma busca. Para validar essa hipótese, foi apresentado um esquema de *ranking* denominado

AuthorityRank (AR), o qual foi implementado em um SRI chamado de *AuthoritySearch*. O esquema AR foi comparado com um esquema tradicional para RI chamado de *tf-idf* e com outro esquema de *ranking* que, assim como o AR, considera o arcabouço estabelecido pela abordagem de *Folkauthority*, o qual foi chamado de AR/*tf-idf*.

Como não existe nenhuma rede social atual que represente uma cadeia de autoridades, foi realizada uma simulação para a definição das autoridades cognitivas no sistema *AuthoritySearch*. Essa simulação foi baseada em uma aproximação, guiada por uma heurística que definia a concessão de autoridades a partir da topologia da rede gerada pela ferramenta *Network* do sistema *Delicious*, bem como das *tags* utilizadas por esses usuários, conforme discutido no Capítulo 4, Seção 4.2. Apesar de se ter empregado uma "aproximação", foi possível observar que a consideração das relações entre os membros de uma rede social na *Web* pode de fato auxiliar na tarefa de RI, sendo que essa afirmação pode ser verificada para o caso de um sistema que utilize a abordagem de *Folkauthority*, conforme demonstram os resultados apresentados no Capítulo 5.

A implementação do esquema de *ranking* AR no sistema *AuthoritySearch* merece ser comentado. Conforme foi discutido no Capítulo 4, o referido esquema de *ranking* exige que seja realizado o cálculo do valor do *PageRank* à Priori de cada nó da cadeia de autoridades, o qual é utilizado como um *boost* para o *score* dos documentos categorizados por autoridades, sendo o cálculo do *PageRank* realizado antes da indexação dos documentos. Esse fato implica na necessidade de se atualizar o índice do sistema *AuthoritySearch* de forma eficiente sempre que uma autoridade seja categorizada, pois o cálculo do *PageRank* considera a propagação de autoridades na cadeia (Pereira e da Silva, 2008b) (isto é, a introdução de uma nova autoridade pode influenciar no valor do *PageRank* de outra autoridade). Por outro lado, realizar o cálculo em tempo de indexação traz vantagens com relação ao tempo computacional de uma consulta, pois o valor do *PageRank* já encontra-se estipulado no momento de se calcular o *score* dos documentos (caso contrário, seria necessário estipular esse valor no momento da consulta, fazendo com que fosse adicionado ao tempo computacional da consulta pelo menos um tempo computacional proporcional à complexidade do cálculo do valor do *PageRank*).

Tendo em vista a hipótese de utilização da abordagem de *Folkauthority* para a tarefa de RI, um dos propósitos dos testes realizados neste trabalho vem de encontro à avaliação dos esquemas de *ranking* citados, de forma que fosse possível verificar qual esquema recuperava informação de melhor relevância e qualidade nas primeiras posições do *rank*. O estudo realizado foi baseado: i) na comparação da eficiência de dois diferentes esquemas de *ranking*, os quais consideram as autoridades cognitivas e, ii) na comparação de dois esquemas: um que não considera as autoridades cognitivas no cálculo do *rank* e outro

que considera. Para realizar a avaliação sobre um ponto de vista quantitativo, foi lançado mão da métrica de NDCG – a qual foi aplicada à um conjunto incompleto e fechado de documentos, julgados quanto à relevância e à qualidade por usuários participantes – e do teste de significância *T de Student* com um grau de significância de 95%.

Com base na observação da média do NDCG-10 das nove consultas avaliadas para o critério de relevância, foi possível verificar que, i) comparando os resultados dos esquemas de *ranking* AR e *tf-idf*, houve uma melhoria na média de NDCG-10 de 54,4% para os cinco primeiros resultados de busca, 54,0% para os dez primeiros, 34,64% para os vinte primeiros e 23,4% para os trinta primeiros, sendo possível concluir que um esquema de *ranking* que considera a abordagem de *Folkauthority* é capaz de apresentar aos usuários participantes resultados de busca mais relevantes e, ii) comparando os esquemas de *ranking* AR e AR/*tf-idf*, não houve uma diferença significativa com relação aos resultados de busca gerados por esses esquemas ($\alpha = 0,05$), isto é, as duas abordagens que consideram autoridades não geraram resultados significativamente diferentes. Em parte, esse resultado pode ser explicado pelo funcionamento do esquema de *ranking* AR/*tf-idf*, o qual é baseado em uma reordenação dos 30 primeiros resultados de busca utilizando o esquema AR. A restrição de se reordenar somente os 30 primeiros resultados de busca limita a diversidade de documentos nos resultados de busca, de forma que os resultados apresentados pelos dois esquemas sejam diferentes somente com relação à ordem em que são apresentados os documentos, sendo os documentos em si os mesmos. Dessa forma, os valores de NDCG-10 desses documentos, mesmo que apresentados em ordem diferente, não foi significativamente diferente entre os esquemas AR e AR/*tf-idf*.

Para o critério de qualidade, considerando as médias de NDCG-10 geradas para esquemas de *ranking* AR e *tf-idf* foi possível verificar uma melhoria significativa de 32,83% para os cinco primeiros resultados de busca, 32,26% para os dez primeiros, 29,01% para os vinte primeiros e 14,86% para os trinta primeiros ($\alpha = 0,05$). Com base nesses resultados, é possível concluir que o esquema de *ranking* AR também ordenou os resultados de busca de forma que os documentos de maior qualidade fossem apresentados nos primeiros resultados, embora para o critério de relevância essa melhoria tenha sido maior.

Apesar da questão sobre a consideração de diversas dimensões da Qualidade da Informação por parte dos usuários participantes ter sido discutida por diversos autores, conforme apresentado no Capítulo 5 (Harter, 1996; Saracevic et al., 1988; Schamber, 1994), nem todas estas dimensões foram avaliadas neste trabalho, pois as informações recuperadas não possuíam dados suficientes para que todas as dimensões fossem avaliadas, fazendo com que o conhecimento dos usuários sobre os documentos em si fosse muitas vezes limitado ao título, ao conteúdo e à URL. Dessa forma, não foi possível realizar uma

verificação sobre todos os aspectos que estavam presentes na avaliação das informações recuperadas no sistema *AuthoritySearch*, sendo inviável apontar conclusões sobre tais aspectos. No entanto, é possível levantar a questão da influência do contexto em que uma *tag* é utilizada sobre a tarefa de RI, uma vez que a *tag* utilizada para a categorização de informações pode ter sido utilizada em um contexto diferente daquele em que ela foi usada como um termo em uma consulta, fazendo com que o significado do termo no momento da consulta seja diferente daquele empregado no momento da categorização[1]. Essa observação é válida tanto para as *tags* aplicadas aos documentos quanto para as *tags* aplicadas às autoridades cognitivas. Dessa forma, pode-se indicar que a caracterização de contexto no momento da atribuição de *tags* poderia ser empregada junto a um mecanismo o qual considere o contexto na fase de RI. Sendo assim, um trabalho futuro está relacionado à modelagem da abordagem de *Folkauthority* de forma que o contexto em que as *tags* foram utilizadas seja também representado, possibilitando que o funcionamento do esquema de *ranking* AR seja baseado também nessa característica.

É possível expandir o estudo realizado neste trabalho no sentido de aumentar o número de tópicos utilizados na avaliação, a qual foi executada em conjunto com a simulação descrita no Capítulo 4 a fim de validar a proposta do esquema de *ranking* AR, além de comparar seu desempenho médio com outros esquemas. O aumento do número de tópicos utilizados na avaliação pode revelar de forma mais clara e significativa a relação entre os tópicos, as *tags* atribuídas às autoridades (isto é, a cadeia de autoridades) e a qualidade/relevância das informações apresentadas pelos esquemas de *ranking* em cada tópico. Por exemplo, seria viável uma análise sobre a diferença entre o número de documentos categorizados pelas autoridades cognitivas em cada tópico. Uma vez calculados os valores do *PageRank* à Priori de cada autoridade e os respectivos valores de NDCG dos documentos categorizados por essas autoridades, é possível verificar se há uma relação entre o aumento do número de autoridades cognitivas (e também suas "importâncias") e o aumento da relevância/qualidade das informações em cada tópico. Assim sendo, pode-se melhor compreender a relação entre o processo de concessão de autoridades cognitivas em um SBF e aspetos da qualidade das informações.

Por fim, o estudo realizado ficou restrito à recuperação de documentos no sistema *AuthoritySearch*, sendo possível também a exploração da recuperação de autoridades cognitivas em determinado(s) assunto(s). Apesar de ser trivial utilizar os algoritmos propostos para se recuperar autoridades ao invés de documentos (pois as autoridades também são categorizadas com *tags*), os resultados apresentados não podem ser generalizados para

[1]Este fenômeno é conhecido como polissemia e refere-se ao fato de um mesmo termo possuir vários significados.

a recuperação de autoridades, uma vez que o julgamento da qualidade de uma informação pode ser diferente do julgamento da "qualidade" de uma autoridade cognitiva. Dessa forma, outro trabalho futuro é a investigação da recuperação de autoridades cognitivas a partir dos algoritmos propostos neste trabalho.

Com o desenvolvimento deste trabalho espera-se contribuir para as discussões em torno do tema de Recuperação Social de Informação, definindo novas formas de utilização do conhecimento gerado pelos usuários para a tarefa de RI no ambiente da *Web 2.0*, possibilitando gerar resultados de busca com maior qualidade e relevância. Este trabalho também apontou melhorias em trabalhos anteriores sobre a RI utilizando a abordagem de *Folkauthority*.

Referências

Abel, F.; Hannover, A. D.; Henze, N.; Krause, D. *Analyzing Ranking Algorithms in Folksonomy Systems.* Relatório Técnico, L3S Research Center, 2008.

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, 2005.

Aho, A. V.; Corasick, M. J. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, v. 18, n. 6, p. 333–340, 1975.

Alexander, J. E.; Tate, M. A. *Web wisdom: How to evaluate and create information quality on the web.* 1st ed. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1999.

Amento, B.; Terveen, L.; Hill, W. Does authority mean quality? predicting expert quality ratings of web documents. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM, 2000, p. 296–303.

Angeletou, S.; Sabou, M.; Specia, L.; Motta, E. Bridging the gap between folksonomies and the semantic web: An experience report. In: *Workshop: Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference*, 2007, p. 93.

Baeza-yates, R.; Castillo, C. Balancing volume, quality and freshness in web crawling. In: *Soft Computing Systems - Design, Management and Applications*, IOS Press, 2002, p. 565–572.

Baeza-Yates, R. A.; Ribeiro-Neto, B. *Modern information retrieval.* Addison-Wesley Longman Publishing Co., Inc., 1999.

Bailey, P.; Craswell, N.; Hawking, D. Engineering a multi-purpose test collection for web retrieval experiments. *Information and Processing Management*, v. 39, p. 853–871, 2003.

Bao, S.; Xue, G.; Wu, X.; Yu, Y.; Fei, B.; Su, Z. Optimizing web search using social annotations. In: *Proceedings of the 16th International Conference on World Wide Web - WWW '07*, New York, New York, USA: ACM Press, 2007, p. 501–510.

Barry, C. L. User-defined relevance criteria: an exploratory study. *Journal of American Society for Information Science*, v. 45, p. 149–159, 1994.

Begelman, G.; Keller, P.; Smadja, F. Automated tag clustering: Improving search and exploration in the tag space. In: *Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland*, 2006, p. 15–33.

Berkhin, P. *Survey of clustering data mining techniques.* Relatório Técnico, Accrue Software, 2002.

Berry, M. W.; Drmac, Z.; Jessup, E. R. Matrices, Vector Spaces and Information Retrieval. *SIAM Review*, v. 41, n. 2, p. 335, 1999.

Borlund, P. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, v. 54, n. 10, p. 913–925, 2003.

Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. In: *Proceedings of the Seventh International Conference on World Wide Web*, Amsterdam, The Netherlands: Elsevier Science Publishers B. V., 1998, p. 107–117.

Buckley, C.; Voorhees, E. M. Retrieval evaluation with incomplete information. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, New York, NY, USA: ACM, 2004, p. 25–32 (*SIGIR '04*,).

Castillo, C. Effective web crawling. *SIGIR Forum*, v. 39, p. 55–56, 2005.

Chen, S.; Zhang, Y. Improve web search ranking with social tagging. In: *Proceedings of the 1st International Workshop on Mining Social Media*, Sevilla, Spain, 2009.

Chevalier, M.; Julien, C.; Soule-Dupuy, C. *Collaborative and social information retrieval and access: Techniques for improved user modeling.* 1st ed. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2008.

Chi, E. H. Information seeking can be social. *Computer*, v. 42, p. 42–46, 2009.

Cleverdon, C. The cranfield tests on index language devices. In: Sparck Jones, K.; Willett, P., eds. *Readings in information retrieval*, Morgan Kaufmann Publishers Inc., p. 47–59, 1997.

Cogo, F.; da Silva, S. Recuperação de informação utilizando o conceito de folkauthority. In: de Computação, S. B., ed. *Proceedings of XVI Brazilian Simposium on Multimedia and Web*, 2010.

Cothey, V. Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, v. 55, n. 14, p. 1228–1238, 2004.
Disponível em `http://dx.doi.org/10.1002/asi.20078`

Crestani, F. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, v. 11, n. 6, p. 453–482, 1997.

Croft, W. B. Combining Approaches to Information Retrieval. p. 1–36, 2000.

Currell, G.; Dowman, A. *Essential Mathematics and Statistics for Science.* John Wiley & Sons, Ltd., 2009.

Eklund, P.; Goodall, P.; Wray, T. Information retrieval and social tagging for digital libraries using formal concept analysis. In: *Proceedings of IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*, IEEE, 2010, p. 1–6.

Evans, B. M.; Chi, E. H. Towards a model of understanding social search. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, New York, NY, USA: ACM, 2008, p. 485–494.

Foley, C.; Smeaton, A. F. Division of labour and sharing of knowledge for synchronous collaborative information retrieval. *Information Processing and Management*, v. 46, p. 762–772, 2010.

Frakes, W.; Baeza-Yates, R. *Information retrieval: Data structures and algorithms.* Prentice Hall, 1992.

Fritch, J. W.; Cromwell, R. L. Evaluating internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of American Society for Information Science and Techcnology*, v. 52, n. 6, p. 499–507, 2001.

FUHR, N. Probabilistic models in information retrieval. *The Computer Journal*, v. 35, p. 243–255, 1992.

FUHR, N. Models in information retrieval. In: *Lectures on information retrieval*, Springer-Verlag New York, Inc., p. 21–50, 2001.

GOFFMAN, W. On relevance as a measure. *Information Storage and Retrieval*, v. 2, n. 3, p. 201 – 203, 1964.

GOH, D.; GOH, D.; FOO, S. *Social information retrieval systems: Emerging technologies and applications for searching the web effectively.* Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2007.

GOLBECK, J.; HENDLER, J. Inferring binary trust relationships in web-based social networks. *ACM Transactions on Internet Technology*, v. 6, p. 497–529, 2006.

GOLDBARG, M. C.; LUNA, H. P. *Otimização combinatória e programação linear.* 2a ed. Rio de Janeiro: Elsevier, 2005.

GOLDER, S.; HUBERMAN, B. *The structure of collaborative tagging systems.* Relatório Técnico, HP Labs, 2006.

GOLOVCHINSKY, G. What the query told the link: the integration of hypertext and information retrieval. In: *Proceedings of the Eighth ACM conference on Hypertext*, New York, NY, USA: ACM, 1997, p. 67–74.

GREISDORF, H. Relevance: An Interdisciplinary and Information Science Perspective. *Informing Science*, v. 3, n. 2, p. 67–71, 2000.

HALPIN, H.; ROBU, V.; SHEPHERD, H. The complex dynamics of collaborative tagging. In: PRESS, A., ed. *Proceedings of International World Wide Web Conference*, Association for Computing Machinery, 2007, p. 211–220.

HARTER, S. P. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, v. 47, p. 37–49, 1996.

HEARST, M. A. *Search user interfaces.* Cambridge University Press, 2009.

HIMMA, K. E. The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition. *Ethics and Information Technology*, v. 9, p. 259–272, 2007.

Hirai, J.; Raghavan, S.; Garcia-Molina, H.; Paepcke, A. Webbase: a repository of web pages. *Computer Networks*, v. 33, n. 1–6, p. 277 – 293, 2000.

Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Information retrieval in folksonomies: Search and ranking. In: *Proceedings of the 3rd European Semantic Web Conference*, Springer, 2006, p. 411–426 (*LNCS*, v.4011).

Janes, J. W. Relevance judgments and the incremental presentation of document representations. *Information Processing and Management*, v. 27, n. 6, p. 629–646, 1991.

Janes, J. W. Other peoples judgments: A comparison of users and others judgments of document relevance, topicality, and utility. *JASIS*, v. 45, n. 3, p. 160–171, 1994.

Jansen, B.; Spink, A.; Bateman, J.; Saracevic, T. Real life information retrieval: a study of user queries on the Web. In: *ACM SIGIR Forum*, ACM, 1998.

Jansen, B. J. The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science*, v. 61, n. 8, p. 1517–1534, 2010.

Järvelin, K.; Kekäläinen, J. Ir evaluation methods for retrieving highly relevant documents. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM, 2000, p. 41–48.

Jarvelin, K.; Kekalainen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, v. 20, n. 4, p. 422–446, 2002.

Jin, Z.; Fine, S. The Effect of Human Behavior on the Design of an Information Retrieval System Interface. *Library and Information Science*, v. 28, p. 249–260, 1996.

Kagolovsky, Y.; Mohr, J. R. A new approach to the concept of relevance in Information Retrieval. *Studies In Health Technology And Informatics*, v. 84, n. 1, p. 348–352, 2001.

Karamuftuoglu, M. Collaborative information retrieval: Toward a social informatics view of ir interaction. *Journal of the American Society for Information Science*, v. 49, n. 12, p. 1070–1080, 1998.

Kargar, M. J. Evaluating Weblog Behavior in Quality of Information Criteria. *International Journal on Internet and Distributed Computing*, v. 1, n. 1, p. 16–21, 2011.

KATERATTANAKUL, P.; SIAU, K. Measuring information quality of web sites: development of an instrument. In: *Proceedings of the 20th international conference on Information Systems*, ICIS '99, Atlanta, GA, USA: Association for Information Systems, 1999, p. 279–285 (*ICIS '99*,).

KAZAI, G.; MILIC-FRAYLING, N. Trust, authority and popularity in social information retrieval. In: *CIKM'08: Proceeding of the 17th ACM Conference on Information and Knowledge Mining*, New York, NY, USA: ACM, 2008, p. 1503–1504.

KEEN, E. M. Evaluation parameters. In: SALTON, G., ed. *The SMART Retrieval System — Experiments in Automatic Document Processing*, 1971.

KIM, S.; KWON, J. Information Retrieval using Context Information on the Web 2.0 Environment. *International Journal of Computer Science and Network Security*, v. 9, n. 10, p. 62–65, 2009.

KIM, Y.-M. Social tags in text and image search. In: *Proceeding of the Third Symposium on Information Interaction in Context*, New York, NY, USA: ACM, 2010, p. 353–358.

KIRSCH, S.; GNASA, M.; WON, M.; CREMERS, A. From pagerank to social rank: Authority based retrieval in information spaces. In: GOH, D.; FOO, S., eds. *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively*, Information Science Reference, p. 134–152, 2008.

KLEINBERG, J. Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, v. 46, n. 5, p. 604–632, 1999.

KNIGHT, S. A.; BURN, J. Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, v. 8, p. 159–172, 2005.

KOME, S. *Hierarchical subject relationships in folksonomies.* Master's, University of North Carolina, 2005.

KRAUSE, B.; JÄSCHKE, R.; HOTHO, A.; STUMME, G. Logsonomy - social information retrieval with logdata. In: *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, New York, NY, USA: ACM, 2008, p. 157–166.

LACHICA, R.; KARABEG, D.; RUDAN, S. Quality, relevance and importance in information retrieval with fuzzy semantic networks. *TMRA Germany*, 2008.

Li, L.; Shang, Y.; Zhang, W. Improvement of hits-based algorithms on web documents. In: *Proceedings of the 11th international conference on World Wide Web*, New York, NY, USA: ACM, 2002, p. 527–535.

Maniu, S.; Cautis, B.; Abdessalem, T. Efficient top-k retrieval in real social tagging networks. *Computing Research Repository - CoRR*, v. abs/1104.1605, 2011.

Manning, C.; Raghavan, p.; Schütze, H. *Introduction to information retrieval*. Cambridge University Press, 2009.

Marchionini, G. Exploratory Search: From finding to understanding. *Communications of the ACM*, v. 49, n. 4, p. 41–46, 2006.

Mathes, A. *Folksonomies: Cooperative classification and communication through shared metadata*. Relatório Técnico, Graduate School of Library and Information Science at University of Illinois Urbana Champaign, 2005.

McCandless, M.; Hatcher, E.; Gospodnetic, O. *Lucene in action*. Manning Publications Co., 2010.

McKenzie, P. Justifying cognitive authority decisions: Discursive strategies of information seekers. *Library Quarterly*, v. 73, n. 3, p. 261–288, 2003.

Melnik, S.; Raghavan, S.; Yang, B.; Garcia-Molina, H. Building a distributed full-text index for the Web. *ACM Transactions on Information Systems*, p. 396–406, 2001.

Melucci, M. A basis for information retrieval in context. *ACM Transactions on Information Systems*, v. 26, n. 3, p. 1–41, 2008.

Metzger, M. J. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, v. 58, p. 2078–2091, 2007.

Mika, P. *Social networks and the semantic web*, v. 5 de *Semantic Web and Beyond*. New York: Springer, 2007.

Murugesan, S. Understanding web 2.0. *IT Professional*, v. 9, p. 34–41, 2007.

O'Reilly, T. What is web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies*, v. 1, n. 1, p. 17–37, 2007.

Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank citation ranking: bringing order to the Web.* Relatório Técnico, Stanford Digital Library Technologies Project, 1998.

Pereira, R. *A aplicação do conceito de autoridade cognitiva por meio de folksonomia.* Dissertaćão de Mestrado, Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Maringá, 2008.

Pereira, R.; da Silva, S. Chain of authorities: Authority and information quality. In: *Workshop on Adaptation for the Social*, 2008a, p. 24–33.

Pereira, R.; da Silva, S. Folksonomias: Uma análise crítica focada na interação e na natureza da técnica. In: de Computação, S. B., ed. *Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems*, Association for Computing Machinery, 2008b, p. 126–135.

Pereira, R.; da Silva, S. R. P. The use of cognitive authority for information retrieval in folksonomy-based systems. In: *Proceedings of the 2008 Eighth International Conference on Web Engineering*, Washington, DC, USA: IEEE Computer Society, 2008c, p. 325–331.

Petratos, P. Information Retrieval Systems: A Perspective on Human Computer Interaction. *Issues in Informing Science and Information Technology*, v. 3, p. 511–518, 2006.

Preece, S. E. *A spreading activation network model for information retrieval.* Tese de Doutoramento, Champaign, IL, USA, 1981.

Primo, A. O aspecto relacional das interações na web 2.0. *Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação*, v. 9, n. 1, p. 1–21, 2007.

Rieh, S. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, v. 53, n. 2, p. 145–161, 2002.

Robertson, S. E.; Jones, K. S. Relevance weighting of search terms. *Journal of the American Society for Information Science*, v. 27, n. 3, p. 129–146, 1976.

Rogers, Y.; Sharp, H.; Preece, J. *Interaction Design: Beyond Human-Computer Interaction.* John Wiley and Sons Ltd, 2002.

ROSCH, E. Principles of categorization. In: ROSCH, E.; LLOYD, B., eds. *Cognition and categorization*, Hillsdale, New Jersey: Erlbaum, p. 27–48, 1978.

RUSSEL, T. *Contextual authority tagging: Cognitive authority through folksonomy.* Relatório Técnico, School of Information and Library Science at South Carolina University, 2005.

SALTON, G. *The smart retrieval system.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

SALTON, G.; FOX, E. A.; WU, H. Extended Boolean information retrieval. *Communications of the ACM*, v. 26, n. 11, p. 1022–1036, 1983.

SANDERSON, M. Test collection based evaluation of information retrieval systemss. *Foundations and Trends in Information Retrieval*, v. 4, p. 247–375, 2010.

SARACEVIC, T. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, v. 26, n. 6, p. 321–343, 1975.

SARACEVIC, T. Relevance: A review of the literature and a framework for thinking on the notion in information science – part ii: nature and manifestations of relevance. *Journal of American Society for Information Science and Technology*, v. 58, p. 1915–1933, 2007.

SARACEVIC, T.; KANTOR, P.; CHAMIS, A. Y.; TRIVISON, D. A study in information seeking and retrieving – i. background and methodology. *Journal of the American Society for Information Science*, p. 176–195, 1988.

SAVOLAINEN, R. Media credibility and cognitive authority: The case of seeking orienting information. *Information Research*, v. 12, n. 3, 2007.

SCHAMBER, L. Relevance and information behavior. *Annual Review of Information Science and Technology*, v. 29, p. 3–48, 1994.

SCHAMBER, L.; EISENBERG, M.; NILAN, M. S. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing and Management*, v. 26, p. 755–776, 1990.

SCHENKEL, R.; CRECELIUS, T.; KACIMI, M.; MICHEL, S.; NEUMANN, T.; PARREIRA, J. X.; WEIKUM, G. Efficient top-k querying over social-tagging networks. In:

Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, New York, NY, USA: ACM, 2008, p. 523–530.

Sen, S.; Lam, S.; Rashid, A.; Cosley, D.; Frankowisk, D.; Osterhouse, J.; Harper, M.; Riedl, J. tagging, communities, vocabulary, evolution. In: Press, A., ed. *Proceedings of Conference on Computer Suported Cooperative Work*, 2006, p. 181–190.

Singhal, A. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, v. 24, n. 4, p. 35–43, 2001.

Smucker, M. D.; Allan, J.; Carterette, B. A comparison of statistical significance tests for information retrieval evaluation. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, New York, NY, USA: ACM, 2007, p. 623–632.

Song, Y.; Zhang, L.; Giles, C. L. A sparse gaussian processes classification framework for fast tag suggestions. In: *Proceeding of the 17th ACM conference on Information and Knowledge Management*, 2008, p. 93–102.

Szomszor, M. N.; Cantador, I.; Alani, H. Correlating user profiles from multiple folksonomies. In: *Proceedings of the Nineteenth ACM conference on Hypertext and Hypermedia - HT 08*, ACM Press, 2008.

Trant, J. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, v. 10, n. 1, 2009.

Turtle, H. R.; Croft, B. A Comparison of Text Retrieval Models. *The Computer Journal*, v. 35, n. 3, p. 279–290, 1992.

Vakkari, P.; Sormunen, E. The influence of relevance levels on the effectiveness of interactive information retrieval. *Journal of the American Society for Information Science*, v. 55, n. 11, p. 963–969, 2004.

Valizadegan, H.; Jin, R.; Zhang, R.; Mao, J. Learning to Rank by Optimizing NDCG Measure. In: *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2000, p. 41–48.

Vander Wal, T. *Folksonomy coinage and definition* http://www.vanderwal.net/folksonomy.html, 2007.

VLAHOVIC, N. Information Retrieval and Information Extraction in Web 2.0 Environment. *International Journal of Computers*, v. 5, n. 1, 2011.

VOSS, J. Tagging, folksonomy and co-renaissance of manual indexing ? In: OSSWALD, A., ed. *10th International Symposium for Information Science*, Cologne: UVK-Verlagsgesellschaft, 2007, p. 243–254.

WANG, J.; DAVISON, D. Explorations in tag suggestion and query expansion. In: *Proceeding of the 2008 ACM workshop on Search in Social Media*, New York, New York, USA: ACM Press, 2008, p. 43.

WANG, R. Y.; STRONG, M. D. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, v. 12, n. 4, p. 5–34, 1996.

WATHEN, C. N.; BURKELL, J. Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, v. 53, n. 2, p. 134–144, 2002.

WEBBER, W. *Measurement in information retrieval evaluation.* Tese de Doutoramento, Department of Computer Science and Software Engineering of The University of Melbourne, 2010.

WHITE, S.; SMYTH, P. Algorithms for estimating relative importance in networks. In: *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, 2003, p. 266–275.

WILSON, P. *Second-hand knowledge: An inquiry into cognitive authority.* Greenwood Press, 1983.

YAO, Y. Y. Measuring retrieval effectiveness based on user preference of documents. *Journal of the American Society for Information Science*, v. 46, p. 133–145, 1995.

YILMAZ, E.; KANOULAS, E.; ASLAM, J. A. A simple and efficient sampling method for estimating ap and ndcg. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM, 2008, p. 603–610.

YONG, D. Enhanced web information retrieval by topic tag mining. *Journal of Convergence Information Technology*, v. 6, n. 4, p. 18–24, 2011.

ZHOU, D.; BRIAN, J.; ZHENG, S.; ZHA, H.; GILES, C. L. Exploring Social Annotations for Information Retrieval. In: *Proceedings of the WWW/Internet Conference: Social Networks and Web 2.0 Track*, 2008, p. 715–724.

ZOBEL, J.; MOFFAT, A. Inverted Files for Text Search Engines. *ACM Computing Surveys*, v. 38, n. 2, p. 1–56, 2006.

Apêndice

A

Resultados de NDCG-10 por Tópicos

As Figuras A.1–A.9 mostram os resultados em termos de NDCG-10 para o critério de *relevância* dos documentos avaliados pelos usuários participantes, enquanto as Figuras A.10–A.18 mostram os resultados para o critério de *qualidade*. Para cada critério, são apresentados os gráficos com os valores de NDCG-10 para os nove tópicos considerados nas pesquisas, os quais são diferenciados pela consulta utilizada.

A.1 Critério de Relevância

Consulta: `ajax api`

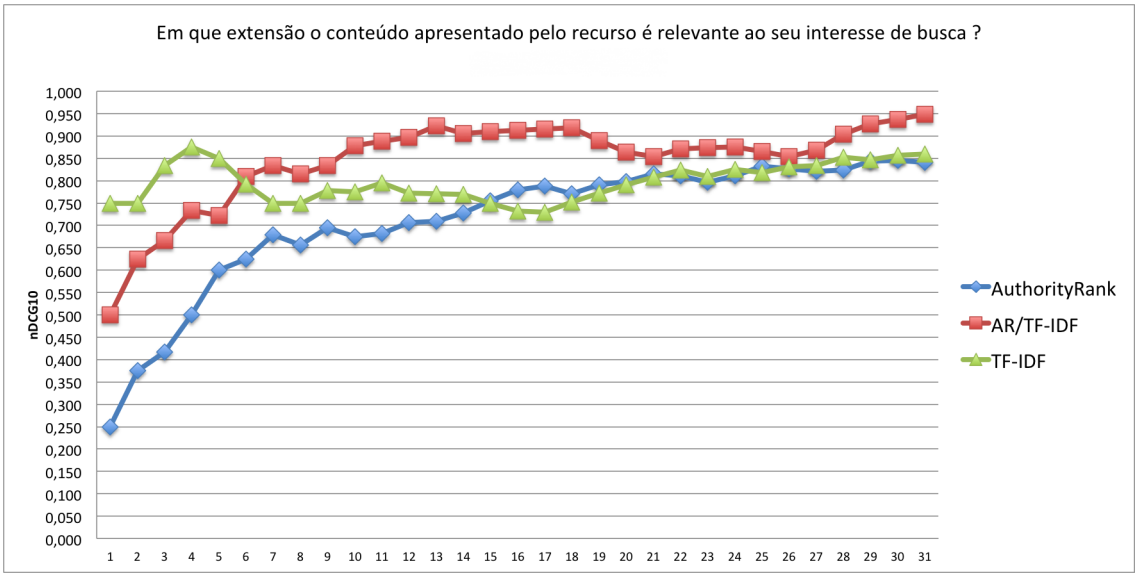

Figura A.1: Resultados de NDCG-10 para a consulta `ajax api` e o critério de relevância.

Consulta: `usability test tutorial`

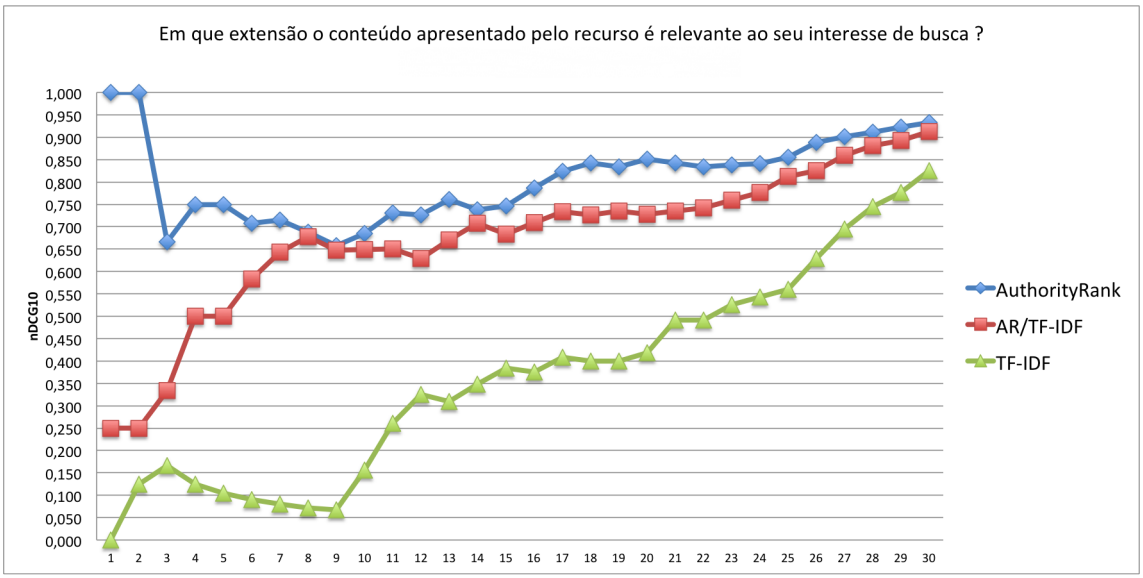

Figura A.2: Resultados de NDCG-10 para a consulta `usability test tutorial` e o critério de relevância.

Consulta: `howto tutorial build make blogs`

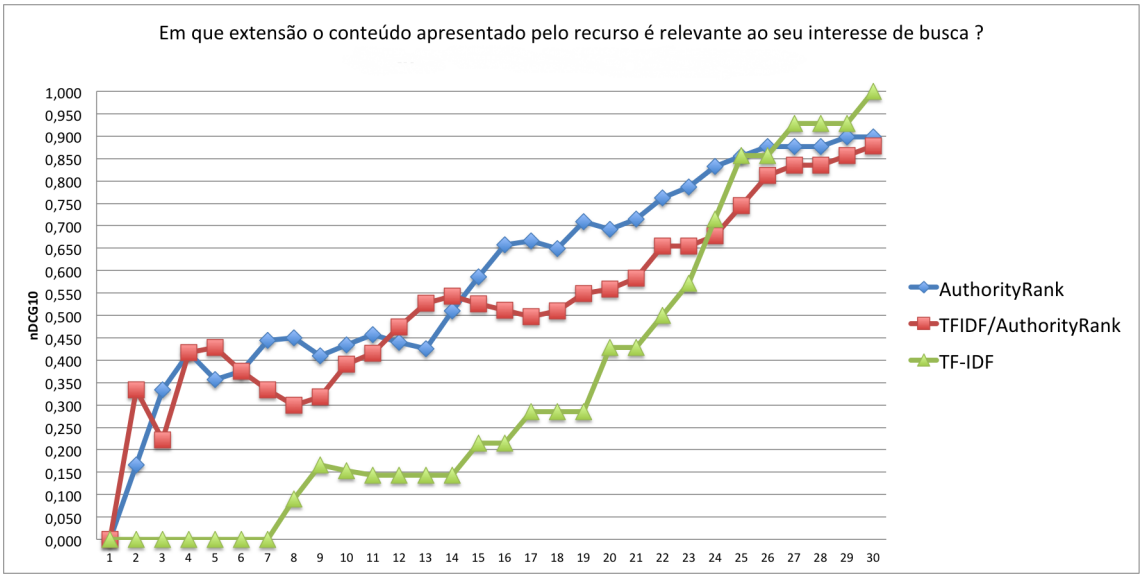

Figura A.3: Resultados de NDCG-10 para a consulta `howto tutorial build make blogs` e o critério de relevância.

Consulta: `interaction design patterns user interface ui`

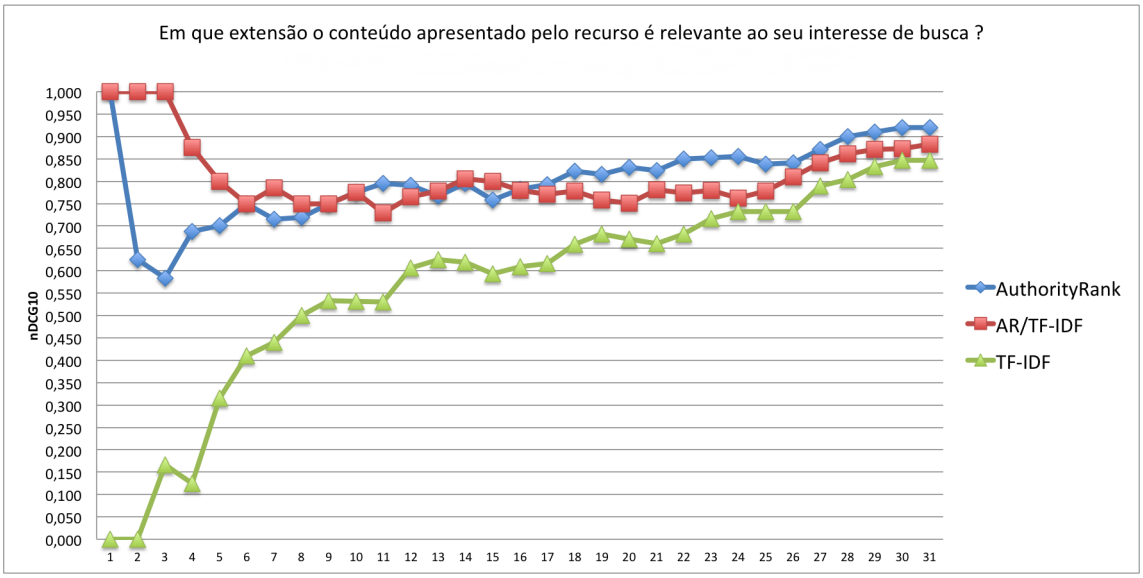

Figura A.4: Resultados de NDCG-10 para a consulta `interaction design patterns user interface ui` e o critério de relevância.

Consulta: `object oriented oo oop uml unified modeling language tool`

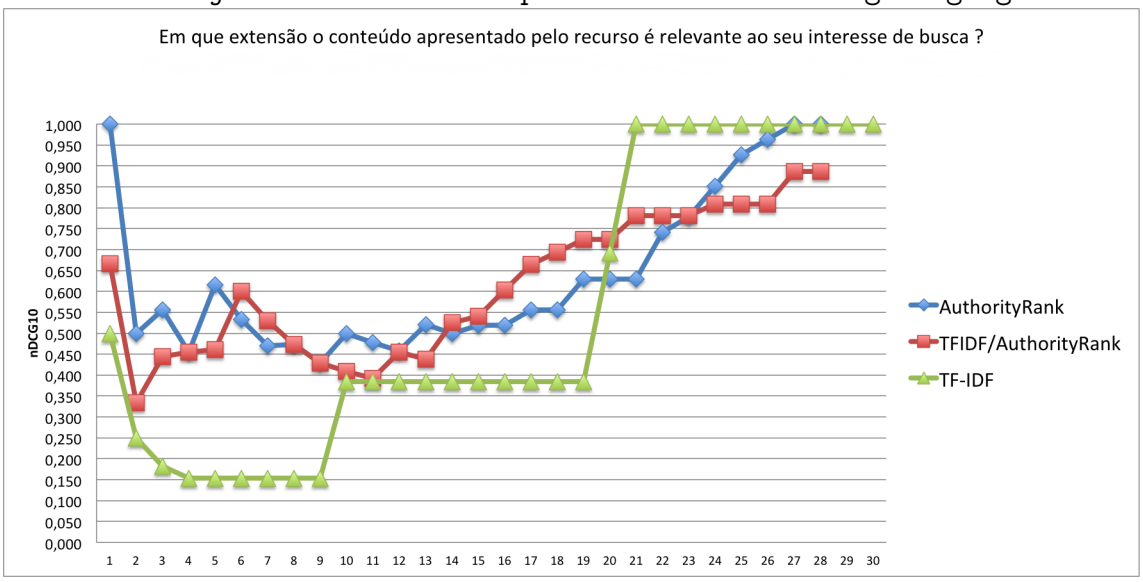

Figura A.5: Resultados de NDCG-10 para a consulta `object oriented oo oop uml unified modeling language tool` e o critério de relevância.

Consulta: `ontology engineering methodology`

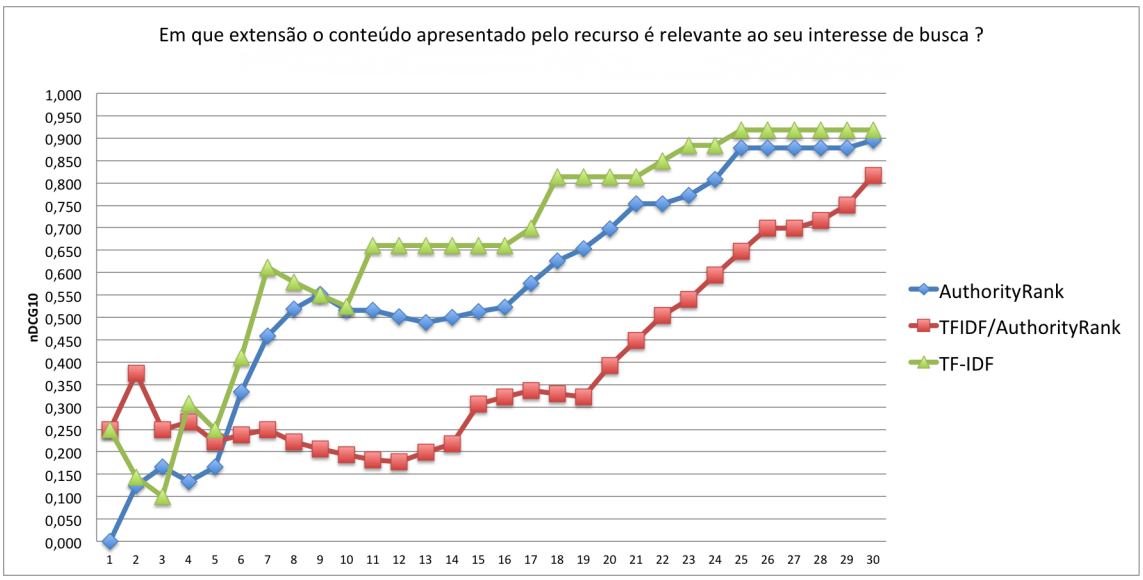

Figura A.6: Resultados de NDCG-10 para a consulta `ontology engineering methodology` e o critério de relevância.

Consulta: `tagging recommendation personomy folksonomy`

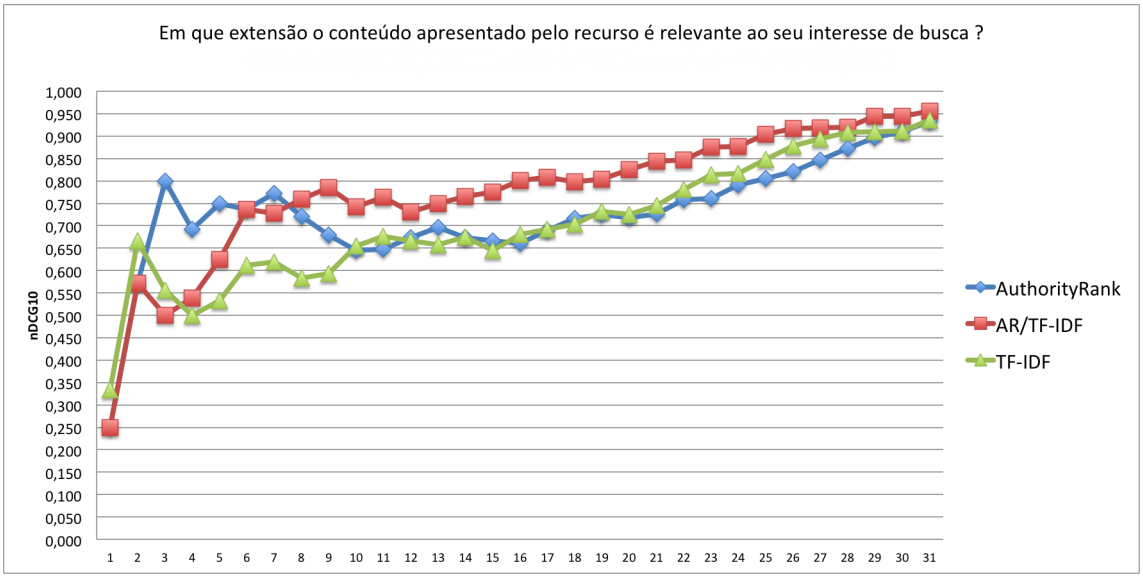

Figura A.7: Resultados de NDCG-10 para a consulta `tagging recommendation personomy folksonomy` e o critério de relevância.

Consulta: `html5 elements description`

[h!]

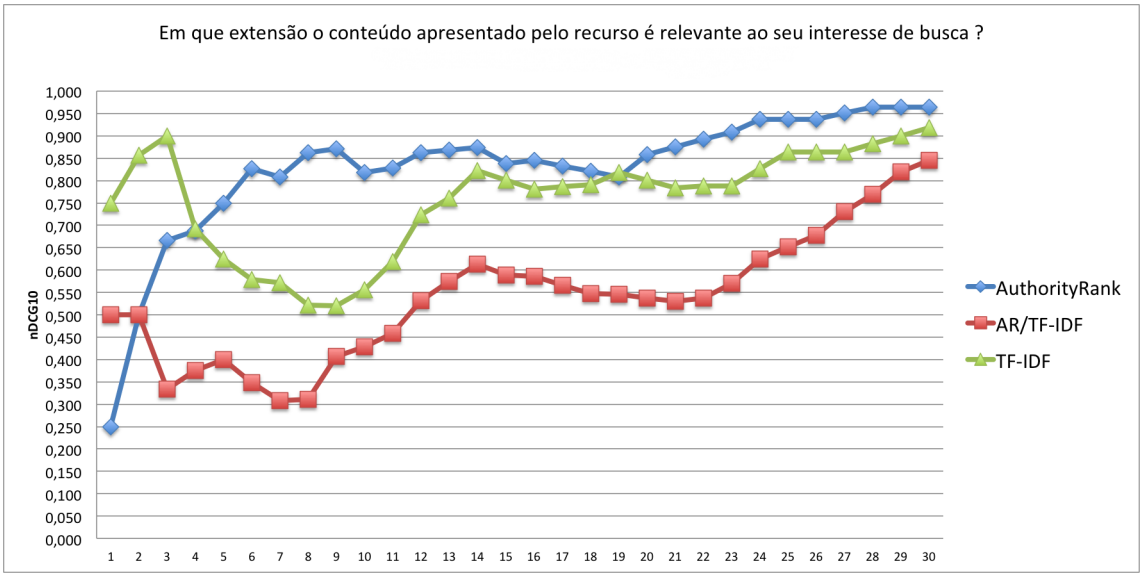

Figura A.8: Resultados de NDCG-10 para a consulta `html5 elements description` e o critério de relevância.

Consulta: `information retrieval visualization`

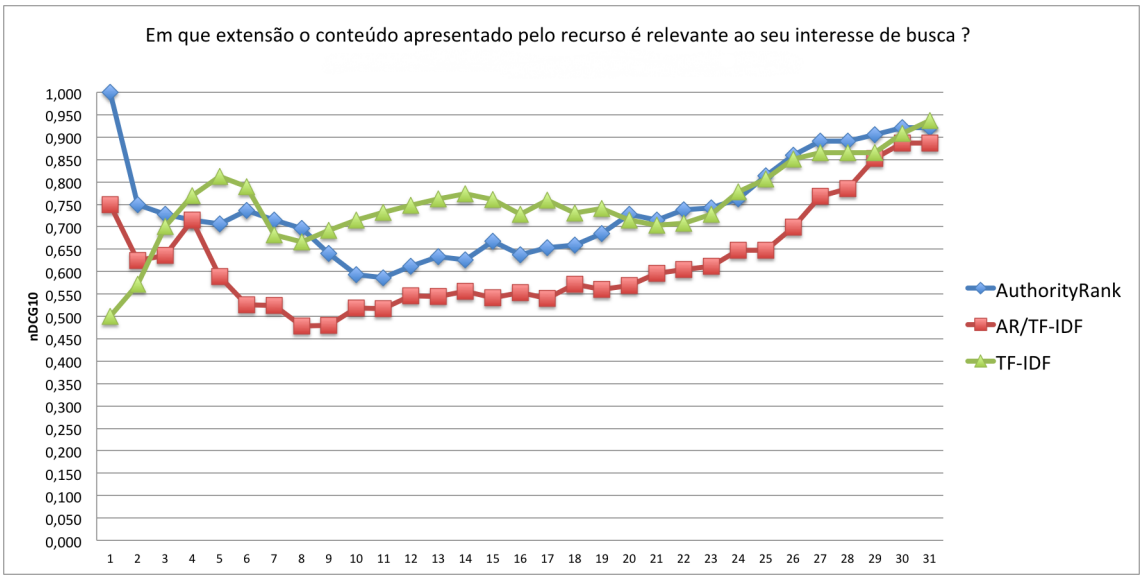

Figura A.9: Resultados de NDCG-10 para a consulta `information retrieval visualization` e o critério de relevância.

A.2 Critério de Qualidade

Consulta: `ajax api`

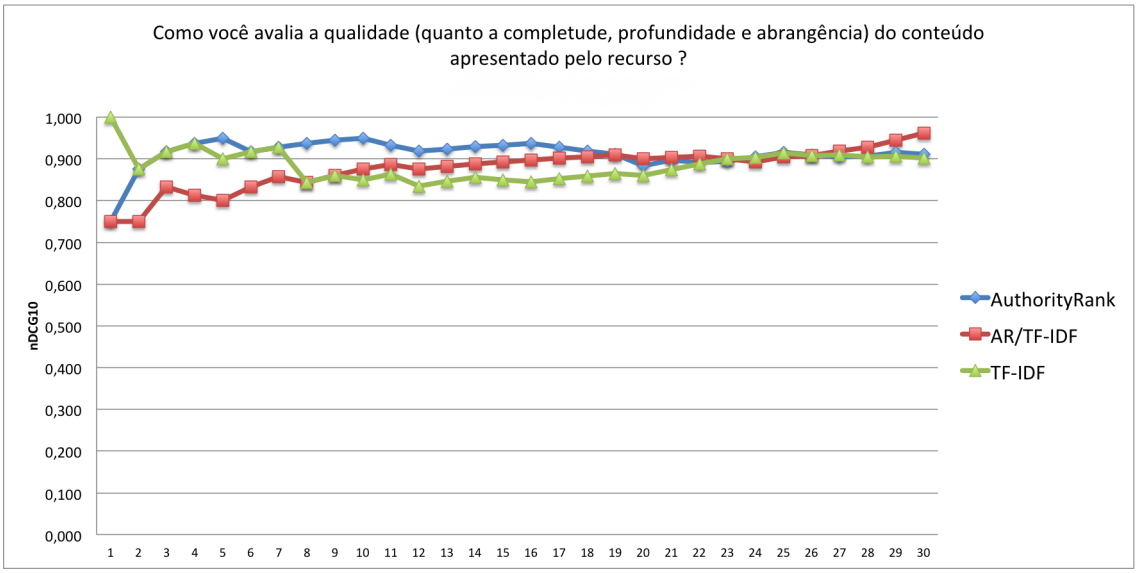

Figura A.10: Resultados de NDCG-10 para a consulta `ajax api` e o critério de qualidade.

Consulta: `usability test tutorial`

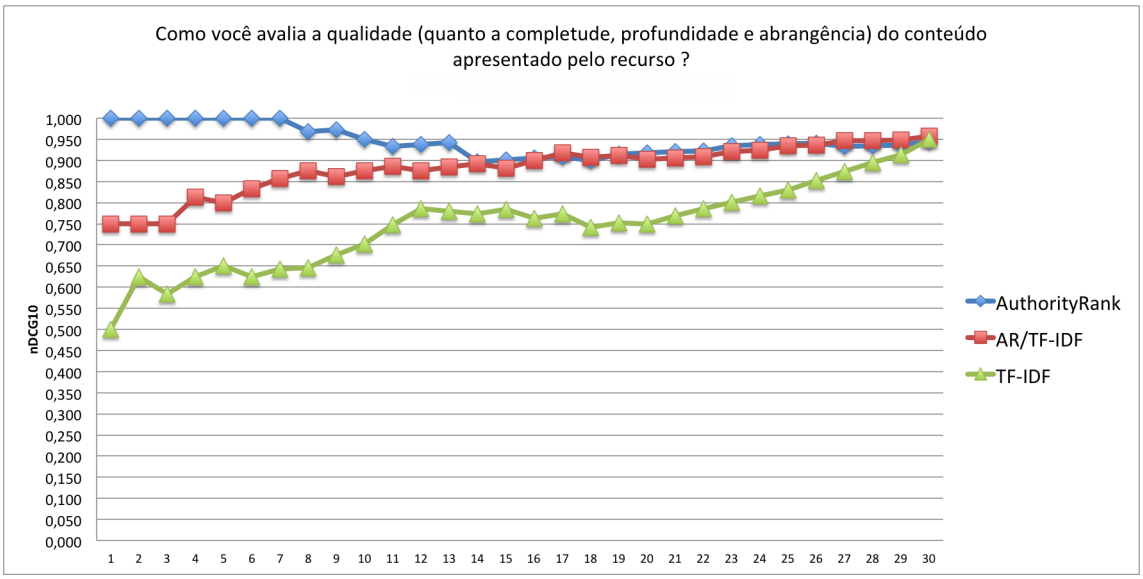

Figura A.11: Resultados de NDCG-10 para a consulta `usability test tutorial` e o critério de qualidade.

Consulta: `howto tutorial build make blogs`

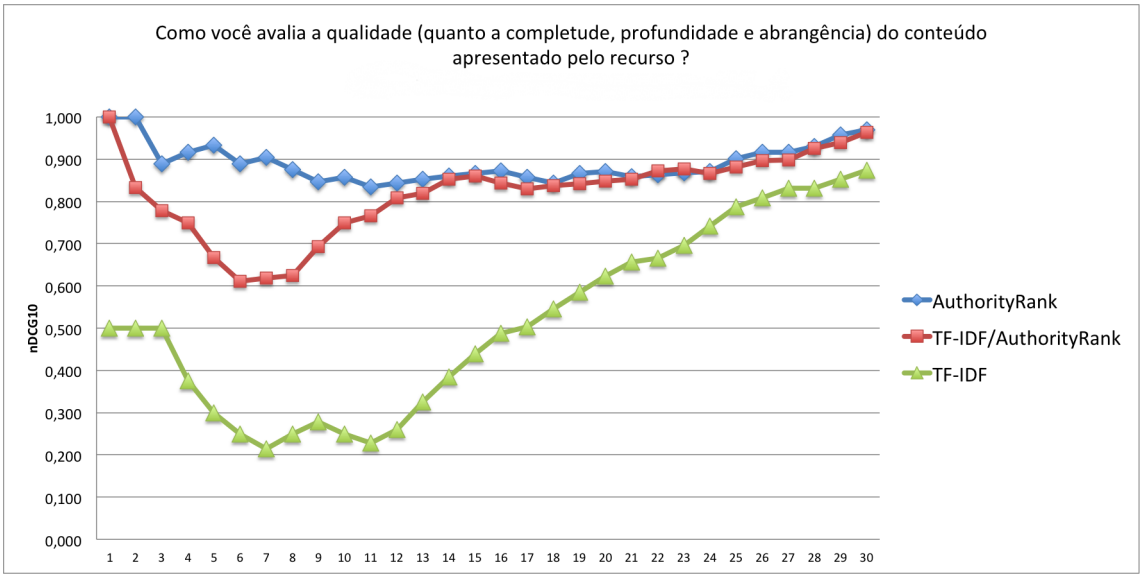

Figura A.12: Resultados de NDCG-10 para a consulta `build blog howto tutorial` e o critério de qualidade.

Consulta: `interaction design patterns user interface ui`

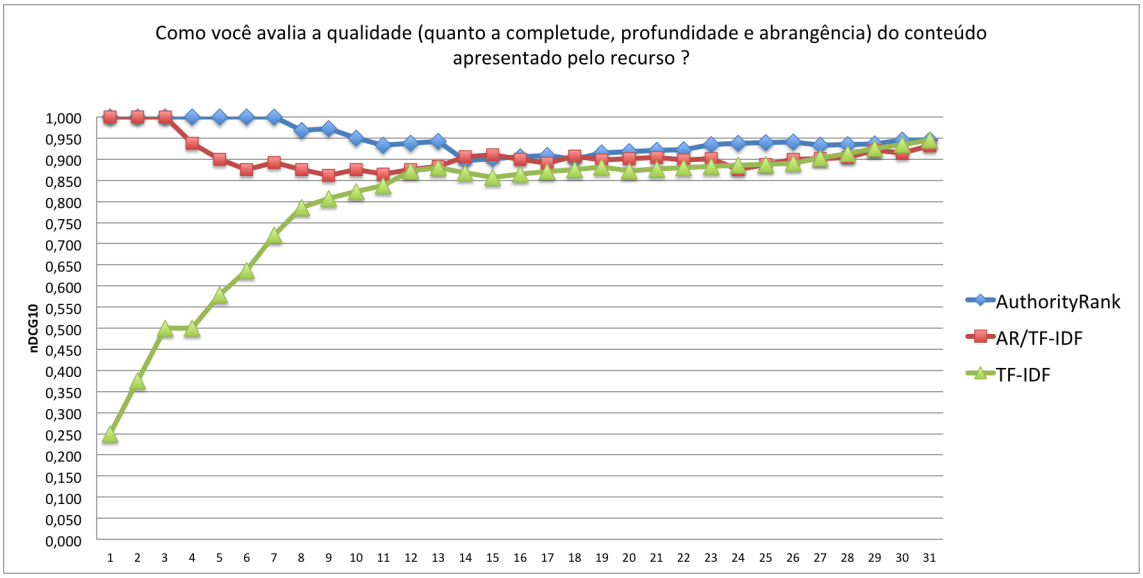

Figura A.13: Resultados de NDCG-10 para a consulta `interaction design patterns user interface ui` e o critério de qualidade.

Consulta: `object oriented oo oop uml unified modeling language tool`

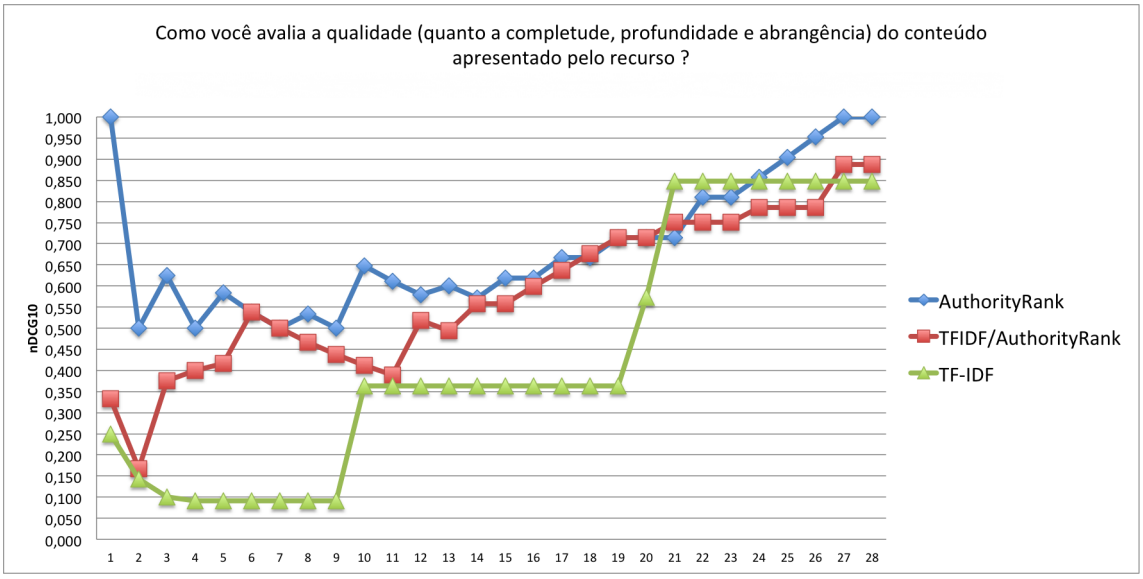

Figura A.14: Resultados de NDCG-10 para a consulta `object oriented oo oop uml unified modeling language tool` e o critério de qualidade.

Consulta: `ontology engineering methodology`

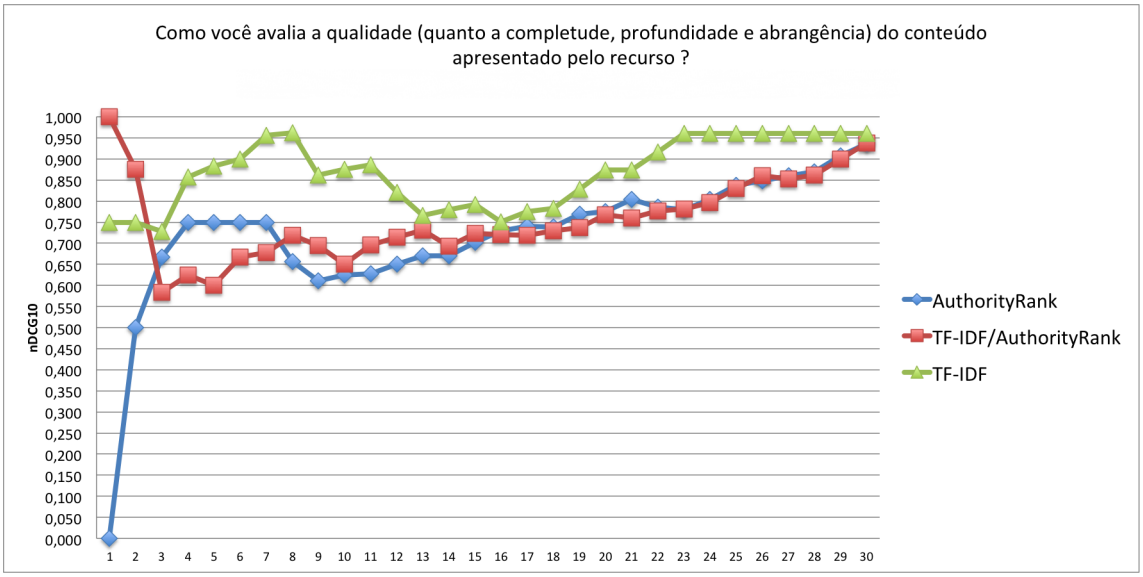

Figura A.15: Resultados de NDCG-10 para a consulta `ontology engineering methodology` e o critério de qualidade.

Consulta: `tagging recommendation personomy folksonomy`

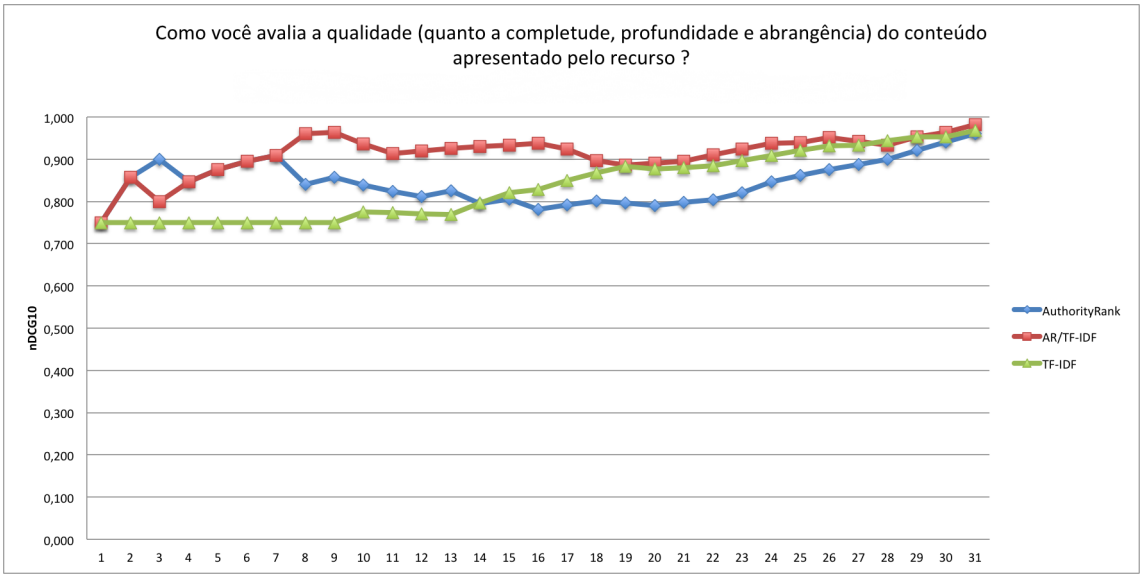

Figura A.16: Resultados de NDCG-10 para a consulta `tagging recommendation personomy folksonomy` e o critério de qualidade.

Consulta: `html5 elements description`

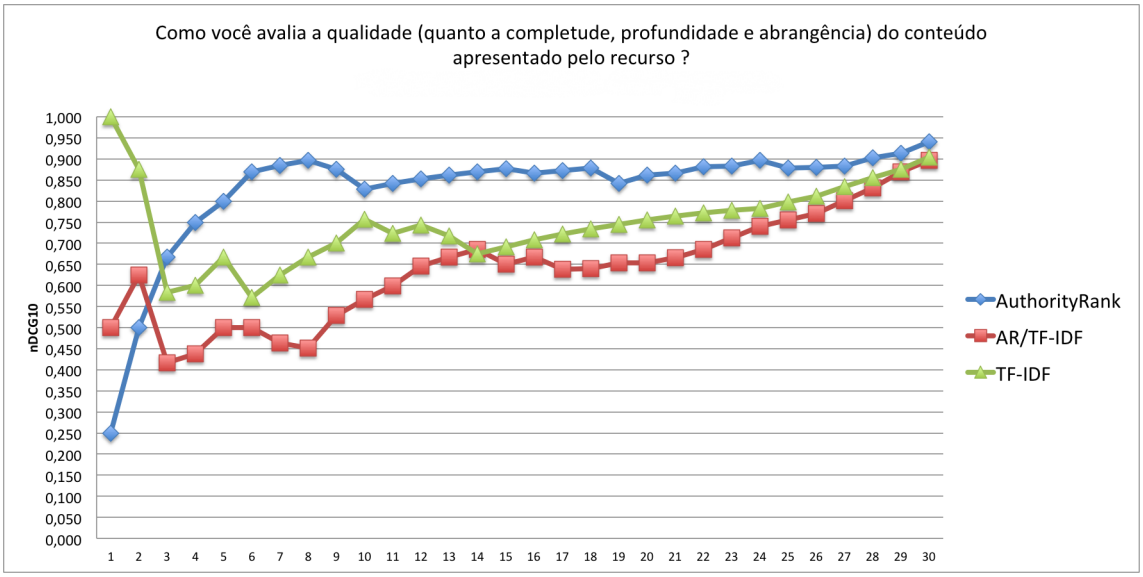

Figura A.17: Resultados de NDCG-10 para a consulta `html5 elements description` e o critério de qualidade.

Consulta: `information retrieval visualization`

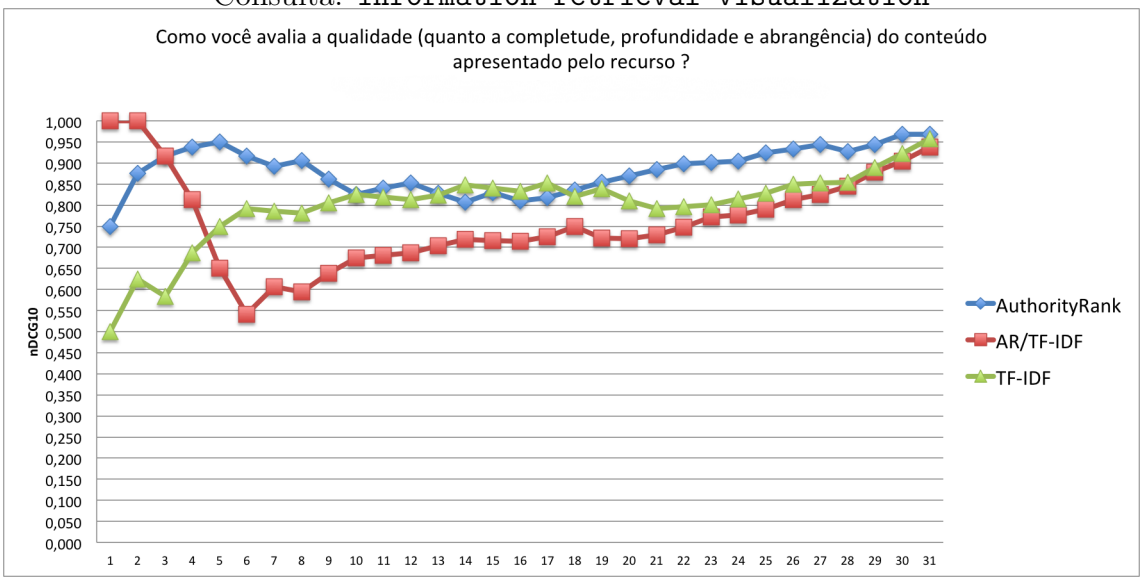

Figura A.18: Resultados de NDCG-10 para a consulta `information retrieval visualization` e o critério de qualidade.

Apêndice

B

Médias de NDCG-10 por faixas de resultados

As Tabelas B.1 e B.2 descrevem os valores de NDCG-10 para os critérios de relevância e qualidade, respectivamente, considerando cada uma das faixas de resultados. A médias dos valores descritos nessas duas tabelas dão origem aos valores descritos nas Tabelas 5.5 e 5.6.

Consulta	***Rank***	**Média por *ranking***		
		AR	AR/TF-IDF	TF-IDF
"ajax api"	Primeiro	0,250	0,500	**0,750**
	Cinco primeiros	0,461	0,676	**0,808**
	Dez primeiros	0,559	0,755	**0,791**
	Vinte primeiros	0,649	**0,822**	0,777
	Trinta primeiros	0,707	**0,842**	0,795
"usability test tutorial"	Primeiro	**1,000**	0,250	0,000
	Cinco primeiros	**0,813**	0,403	0,169
	Dez primeiros	**0,759**	0,517	0,113
	Vinte primeiros	**0,773**	0,600	0,231
	Trinta primeiros	**0,807**	0,673	0,363
"howto tutorial build blog"	Primeiro	0,000	0,000	0,000
	Cinco primeiros	0,275	**0,296**	0,000
	Dez primeiros	**0,350**	0,321	0,050
	Vinte primeiros	**0,459**	0,412	0,128
	Trinta primeiros	**0,585**	0,525	0,292
"interaction design patterns user interface UI"	Primeiro	**1,000**	0,750	0,000
	Cinco primeiros	0,724	**0,736**	0,169
	Dez primeiros	0,797	**0,827**	0,323
	Vinte primeiros	0,759	**0,848**	0,462
	Trinta primeiros	**0,956**	0,880	0,820
"object oriented unified modelling language tools"	Primeiro	**1,000**	0,667	0,500
	Cinco primeiros	**0,610**	0,493	0,232
	Dez primeiros	**0,546**	0,472	0,239
	Vinte primeiros	**0,541**	0,528	0,316
	Trinta primeiros	0,626	**0,629**	0,495
"ontology ontologies engineering method methodology"	Primeiro	0,000	0,250	0,250
	Cinco primeiros	0,154	**0,267**	0,244
	Dez primeiros	0,317	0,242	**0,399**
	Vinte primeiros	0,428	0,263	**0,542**
	Trinta primeiros	0,564	0,389	**0,659**
"tagging recommendation personomy folksonomy"	Primeiro	0,250	0,250	0,333
	Cinco primeiros	0,613	0,497	0,518
	Dez primeiros	0,662	0,624	0,565
	Vinte primeiros	0,674	0,703	0,625
	Trinta primeiros	0,722	0,768	0,700
"html5 elements description"	Primeiro	0,250	0,500	0,750
	Cinco primeiros	0,571	0,422	0,765
	Dez primeiros	0,704	0,391	0,657
	Vinte primeiros	0,774	0,473	0,714
	Trinta primeiros	0,827	0,541	0,758
"information retrieval visualization"	Primeiro	1,000	0,750	0,500
	Cinco primeiros	0,779	0,663	0,671
	Dez primeiros	0,728	0,584	0,690
	Vinte primeiros	0,688	0,567	0,717
	Trinta primeiros	0,733	0,615	0,748

Tabela B.1: Valores de NDCG-10 para os n primeiros resultados de busca em cada consulta, considerando o critério de relevância dos documentos

Consulta	***Rank***	**Média por *ranking***		
		AR	AR/TF-IDF	TF-IDF
"ajax api"	Primeiro	0,750	0,750	**1,000**
	Cinco primeiros	0,891	0,797	**0,924**
	Dez primeiros	**0,913**	0,827	0,902
	Vinte primeiros	**0,916**	0,858	0,883
	Trinta primeiros	**0,912**	0,877	0,888
"usability test tutorial"	Primeiro	**1,000**	0,750	0,500
	Cinco primeiros	**0,914**	0,783	0,601
	Dez primeiros	**0,891**	0,823	0,639
	Vinte primeiros	**0,876**	0,856	0,696
	Trinta primeiros	**0,886**	0,882	0,747
"howto tutorial build blog"	Primeiro	1,000	1,000	0,500
	Cinco primeiros	**0,938**	0,773	0,404
	Dez primeiros	**0,904**	0,736	0,331
	Vinte primeiros	**0,884**	0,782	0,390
	Trinta primeiros	**0,891**	0,820	0,518
"interaction design patterns user interface UI"	Primeiro	1,000	1,000	0,250
	Cinco primeiros	**1,000**	0,952	0,473
	Dez primeiros	**0,984**	0,916	0,619
	Vinte primeiros	**0,952**	0,907	0,733
	Trinta primeiros	**0,946**	0,905	0,788
"ontology ontologies engineering method methodology"	Primeiro	0,000	**1,000**	0,750
	Cinco primeiros	0,569	0,725	**0,811**
	Dez primeiros	0,608	0,708	**0,855**
	Vinte primeiros	0,657	0,716	**0,829**
	Trinta primeiros	0,719	0,756	**0,868**
"object oriented unified modelling language tools"	Primeiro	**1,000**	0,333	0,250
	Cinco primeiros	**0,624**	0,372	0,128
	Dez primeiros	**0,595**	0,403	0,161
	Vinte primeiros	**0,612**	0,495	0,262
	Trinta primeiros	**0,686**	0,602	0,458
"tagging recommendation personomy folksonomy"	Primeiro	0,750	0,750	0,750
	Cinco primeiros	0,846	0,826	0,750
	Dez primeiros	0,857	,879	0,753
	Vinte primeiros	0,830	0,897	0,788
	Trinta primeiros	0,842	0,910	0,832
"html5 elements description"	Primeiro	0,250	0,500	1,000
	Cinco primeiros	0,593	0,496	0,745
	Dez primeiros	0,732	0,499	0,705
	Vinte primeiros	0,797	0,575	0,713
	Trinta primeiros	0,829	0,641	0,749
"information retrieval visualization"	Primeiro	0,750	1,000	0,500
	Cinco primeiros	0,886	0,876	0,629
	Dez primeiros	0,883	0,744	0,714
	Vinte primeiros	0,859	0,729	0,772
	Trinta primeiros	0,880	0,755	0,794

Tabela B.2: Valores de NDCG-10 para os n primeiros resultados de busca em cada consulta, considerando o critério de qualidade dos documentos

Apêndice

C

Testes de hipótese

A verificação da avaliação de um SRI é geralmente realizada com um teste de hipótese estatístico, o qual utiliza os vetores pertencentes ao conjunto O e informações sobre a média e o desvio padrão dos dados presentes nos vetores (ver Capítulo 5). Mais especificamente, neste trabalho foi utilizado o teste *T de Student*, denotado por t_{STAT}. Para este trabalho foi escolhido o teste t de duas amostras e unicaudal. As Tabelas C.1 e C.2 apresentam os dados obtidos com os testes. Além disso, a Tabela C.3 resume os valores mínimos de t_{STAT} para a rejeição da hipótese nula com os respectivos níveis de significância α. Os valores mostrados nessa última tabela são chamados de t_{CRIT}.

C.1 Critério de Relevância

	Valores de t_{STAT} – Critério de relevância					
	AR	*AR/TF-IDF*	*AR*	*TF-IDF*	*AR/TF-IDF*	*TF-IDF*
Cinco primeiros	2,8024		11,1738		6,7782	
Dez primeiros	2,74920		8,67440		6,2148	
Vinte primeiros	**1,6896**		3,7888		2,3826	
Trinta primeiros	**0,9412**		**1,8259**		**1,0272**	

Tabela C.1: Dados do teste de significância t_{STAT} para o critério de relevância dos documentos

C.2 Critério de Qualidade

	Valores de t_{STAT} – Critério de relevância					
	AR	*AR/TF-IDF*	*AR*	*TF-IDF*	*AR/TF-IDF*	*TF-IDF*
Cinco primeiros	3,0230		9,91610		7,3774	
Dez primeiros	4,9290		11,51570		7,8412	
Vinte primeiros	2,8137		7,2724		4,6033	
Trinta primeiros	**1,32955**		2,5247		**1,4840**	

Tabela C.2: Dados do teste de significância t_{STAT} para o critério de qualidade dos documentos

C.3 Valores críticos de t

g.l. \ α	0,100	0,050	0,025	0,010
18	1,7341	2,1009	2,4450	2,8784
38	1,6860	2,0244	2,3337	2,7116
58	1,6716	2,0017	2,3011	2,6633

Tabela C.3: Valores de T_{CRIT} para os diferentes graus de liberdade e intervalos de confiança

Apêndice

D

Roteiro de avaliação dos resultados de busca

Neste apêndice é apresentado o documento utilizado como roteiro de avaliação dos resultados de busca, o qual foi dado a cada um dos usuários participantes da pesquisa no momento de avaliar os documentos. A consulta a esse documento era livre enquanto o participante realizava a atividade de avaliação, de forma que os critérios e os procedimentos utilizados pudessem ser esclarecidos na medida em que o participante sentisse necessidade. Adiante está uma instância do modelo de documento entregue aos usuários participantes.

Roteiro para Avaliação de Resultados de Busca

Sistema FolkauthoritySearch

Introdução

Este documento descreve o roteiro para realização da avaliação de resultados de busca no sistema denominado de FolkauthoritySearch. A comparação e a avaliação de abordagens para a recuperação de informação depende de uma base de documentos julgados como relevantes ou irrelevantes, dada uma consulta. Dessa forma, a avaliação tem o intuito de dar subsídios ao estudo do algoritmo de ordenação no referido sistema, chamado de AuthorityRank.

As informações fornecidas por você (avaliador) serão utilizadas para medir quantitativamente a qualidade e a relevância dos resultados de busca no sistema FokauthoritySearch. Mais especificamente, um estudo comparativo entre os algoritmos de ordenação TF-IDF e AuthorityRank será realizado com base nos dados informados nesta avaliação.

Objetivos e Justificativa

Conforme descrito, a fim de avaliar uma abordagem para recuperação de informação é necessário que se julgue alguns critérios com relação aos documentos da base de documentos passíveis de serem recuperados. Dessa forma, a fim de estimar o valor do algoritmo AuthorityRank para a ordenação de documentos, faz-se necessário uma avaliação dos resultados de algumas consultas utilizando esse algoritmo.

Em diferentes trabalhos foram discutidos critérios utilizados pelos usuários para julgar os documentos recuperados. Nesta pesquisa foram utilizados critérios relacionados a relevância e a qualidade dos documentos, os quais serão descritos na sessão de materiais.

Os objetivos específicos desta pesquisa, bem como da avaliação dos recursos, são descritos a seguir:

1. Determinar, com base na opinião do avaliador e em consultas pré-determinadas, o valor relacionado a cada um dos critérios estabelecidos sobre os documentos.
2. Estimar a relevância e a qualidade de cada um dos documentos recuperados utilizando a opinião do avaliador.
3. Utilizar os documentos avaliados para comparar aspectos da recuperação de informação para cada um dos algoritmos de ordenação utilizados (TF-IDF e AuthorityRank).

Materiais

O sistema FolkauthoritySearch oferece uma interface de busca típica de máquinas de busca Web, na qual utiliza-se um campo de formulário para informar uma consulta. Além disso, nessa interface há a opção de selecionar o algoritmo para a ordenação dos documentos (Busca 1 e Busca 2). Essa interface está demonstrada na Figura I.

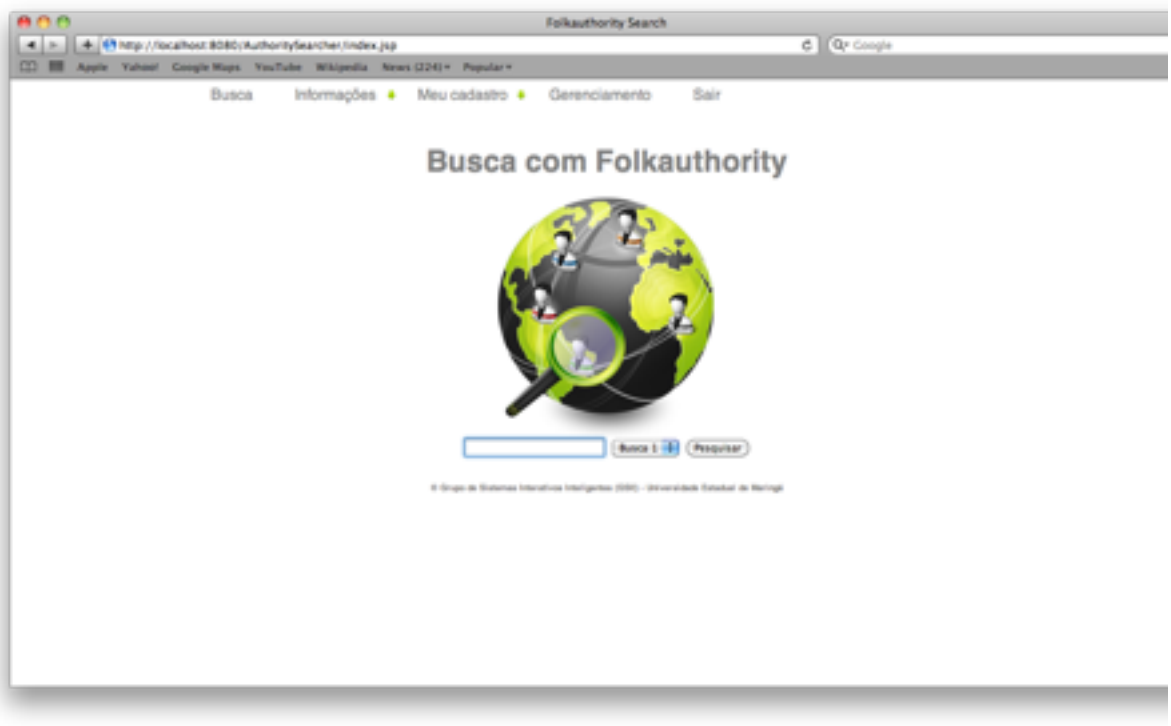

Figura I. Interface de busca do sistema AuthoritySearch

O procedimento para avaliação dos recursos é baseado em uma consulta, a qual deve ser executada utilizando os dois algoritmos de ordenação (busca "Verde" e busca "Amarela"). A consulta, relacionada com o assunto de elementos da linguagem HTML5, é dada pela string "html5 elements description". A relação entre a consulta e a necessidade de informação está descrita na Tabela I:

# Consulta	Consulta	Necessidade de informação
Consulta 1	html5 elements description	Informação sobre a descrição e exemplos de utilização dos elementos da linguagem HTML5.

Tabela I. Consultas e tópicos da necessidade de informação

Essas consultas produzem uma lista ordenada de documentos, os quais devem ser avaliados segundo três critérios diferentes. Cada um desses critérios possui uma escala quintenária que deve ser aferida de acordo com a sua opinião. Os critérios e as escalas estão descritos na Tabela II:

Descrição dos critérios	Escala
Em que extensão o conteúdo apresentado pelo recurso é relevante ao seu interesse de busca ?	Nenhuma (0), Pouca (1), Média (2), Bastante (3), Total (4)
Em que extensão a fonte do conteúdo apresentado (autor, editor, revista, periódico, simpósio, site, etc.) é novo ao seu conhecimento ?	Nenhuma (0), Pouca (1), Média (2), Bastante (3), Total (4)

Descrição dos critérios	Escala
Como você avalia a qualidade (quanto a completude, profundidade e abrangência) do conteúdo apresentado pelo recurso ?	Péssima (0), Ruim (1), Mediana (2), Boa (3), Excelente (4)

Tabela II: Descrição dos critérios x escalas

As noções de "relevância" e "qualidade" do conteúdo são resumidas de acordo com as seguintes definições:

Relevância é a medida na qual a informação contida em um documento é útil para a resolução do(s) problema(s) elencado(s) no cenário de busca proposto. A qualidade resume quanto o conteúdo de um documento é capaz de atender uma necessidade de informação.

Qualidade é a medida na qual o participante julga que a informação contida em um documento seja bem apresentada, verdadeira, abrangente, profunda e completa, sendo considerado um contexto mais amplo para a realização de tal julgamento.

Isto é, um documento pode ser altamente relevante, no sentido de prover informações exatamente sobre o tópico sendo buscado (coluna da necessidade de informação, na Tabela I), no entanto essas informações podem não possuir indicativos de qualidade, tais como boa apresentação, abrangência e profundidade com relação ao assunto ou até mesmo veracidade. Da mesma forma, um documento que contenha informações não tão precisas sobre o tópico sendo buscado, mas que contenha informações sobre esse tópico em um contexto mais amplo, pode possuir altos indicativos de qualidade. Por exemplo, em uma busca cuja necessidade de informação seja relacionada com metodologias de desenvolvimento de software orientado a objetos, um documento que trate sobre engenharia de software ou sobre orientação objetos de forma genérica pode não ser tão relevante, mas pode possuir altos indicativos de qualidade.

Sugere-se também que cada valor da escala seja aferido de acordo com as seguintes premissas, descritas na tabela III:

Grau da escala	Premissas
0	Documento irrelevante, não contém nenhuma informação sobre o tópico relacionado ou Documento de péssima qualidade, não possui nenhum indicativo de aspectos de qualidade.
1	Pouco relevante, o documento apenas aponta para o tópico. Não satisfaz nenhum tópico relacionado à necessidade de informação. ou Baixa qualidade, o documento possui uma apresentação indesejável, conteúdo pouco abrangente ou possui fontes duvidosas.

Grau da escala	Premissas
2	Marginalmente relevante, o documento contém alguma informação sobre o tópico, no entanto não se esgota em nenhuma das informações. Satisfaz apenas parte dos tópicos relacionados à necessidade de informação. ou Média qualidade, o documento possui alguns indicativos de qualidade, tais como uma apresentação desejável ou uma fonte confiável, no entanto não preenche mais do que dois indicativos de qualidade.
3	Relevante, satisfaz todos os tópicos relacionados à necessidade de informação no entanto não é exaustivo nas informações sobre esses tópicos. ou Boa qualidade, o documento possui mais de dois indicativos de qualidade, possuindo informação clara e abrangente sobre o assunto de que trata.
4	Altamente relevante, satisfaz todos os tópicos relacionados à necessidade de informação e apresenta conceitos e exemplos sobre esses tópicos. ou Excelente qualidade, o documento possui todos os indicativos de qualidade citados.

Tabela III: Sugestão de premissas para avaliação dos critérios de relevância e qualidade

A seguir, descreve-se os passos a serem seguidos para realizar a avaliação dos resultados das consultas 1 e 2.

Procedimentos

Os passos descritos nessa sessão referem-se aos procedimentos para: 1) realizar o cadastro no sistema, 2) realizar a indexação dos dados no sistema e, 3) realizar a avaliação dos resultados de busca no sistema.

1. Cadastro no sistema.
   - Na página inicial do sistema, selecione a opção "Cadastrar" no menu superior. Informe seus dados e marque a opção "desejo participar como avaliador do sistema".
   - Clique em "submeter".
2. Indexação dos dados no sistema (somente administrador).
   - Certifique-se que você possui privilégios de administrador e de que esteja logado para realizar os passos do Item 2.

- Selecione a opção "Gerenciamento" no menu superior da página inicial. Selecione a opção "Iniciar a indexação".

3. A valiação dos resultados de busca no sistema.

- Cerfique-se que você possui privilégios de avaliador do sistema. Para tanto, logue no sistema e selecione a opção "Meu cadastro" no menu superior da página inicial. Selecione a opção "desejo participar como avaliador do sistema".

A. Na página incial do sistema (Figura 1), selecione a opção busca "Verde" e realize a Consulta 1.

B. Na página que apresenta os resultados da consulta, para os 30 primeiros documentos apresentados, selecione o ícone para abrir o formulário de avaliação do documento. Esse formulário é mostrado na Figura 2. Ao fim da avaliação clique no botão "submeter". Caso o formulário tenha sido submetido com sucesso, o ícone deverá estar na cor verde . Se algum entre os 30 primeiros recursos estiver indisponível, desconsidere esse recurso na contagem.

C. Realize os mesmos passos utilizando a busca "Amarela".

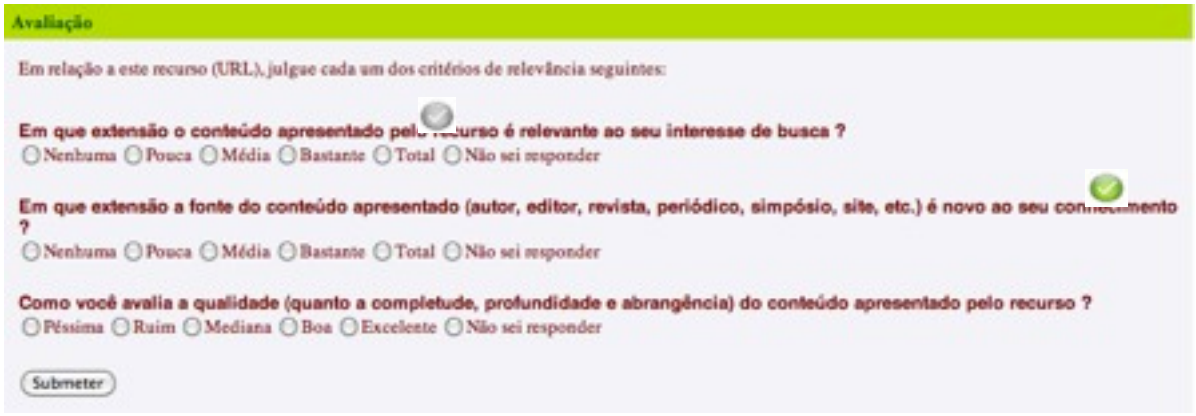

Figura 2: Formulário de avaliação dos recursos.

Agradecemos enormemente a sua ajuda e participação na nossa pesquisa !