

O conceito de relevância e o feedback do usuário final do Repositório Institucional da Universidade de Brasília

The concept of relevance and End user feedback of Institutional Repository of the University of Brasilia

Carlos Duarte Oliveira Jr¹

Leonardo Neves Moreira²

Thalita Franco dos Santos³

Georgia Fernandes do Nascimento⁴

Resumo: Apresenta estudo sobre a relevância adotada pelo Repositório Institucional da Universidade de Brasília e o feedback do usuário às respostas de recuperação da informação. Contextualiza o estudo no universo da comunicação científica e o livre acesso à informação. Identifica conceitos de relevância e abordagens de feedback de relevância para propor ao sistema uma busca que integre tanto pré-requisitos do sistema, quanto à avaliação do usuário, pois este tem papel importante sobre a avaliação por relevância.

Palavras-chave: acesso livre; informação científica; filtragem colaborativa; filtragem de conteúdo; relevância de tópico; relevância de usuário;

Título: El concepto de relevancia y o feedback de los usuarios finales del Repositorio Institucional de la Universidad de Brasilia.

Resumen: Presenta un estudio sobre la relevancia adoptada por el Repositorio Institucional de la Universidad de Brasilia y comentarios de los usuarios a las respuestas de recuperación de información. Contextualiza el estudio dentro del universo de la comunicación científica y el libre acceso a la información. Identifica los conceptos de relevancia y relevancia comentarios acerca de proponer un sistema de búsqueda que integra tanto los requisitos previos del sistema, como la evaluación de usuario, ya que tiene un papel importante en la evaluación por relevancia.

Palabras-clave: acceso abierto, la información científica, filtrado colaborativo, filtrado de contenidos, la pertinencia del tema, la relevancia para el usuario;

Title: The concept of relevance and End user feedback of Institutional Repository of the University of Brasilia

Abstract: Presents a study on the relevance adopted by the Institutional Repository of the University of Brasilia and user feedback to the answers of information retrieval. Contextualizes the study within the universe of scholarly communication and open access to information. Identifies concepts of relevance and relevance feedback approaches to propose a search system that integrates both prerequisites of the system, as the user evaluation, because it has an important role on the assessment by relevance.

Keywords: open access, scientific information, collaborative filtering, content filtering, relevance of topic, relevance to user

¹ Mestrando em Ciência da Informação no Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília – UNB. E-mail: caduartejr@gmail.com

² Mestrando em Ciência da Informação no Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília - UNB. E-mail: leonardolnm@yahoo.com.br

³ Mestre em Ciência da Informação no Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília - UNB. E-mail: thalitafdsantos@gmail.com

⁴ Mestranda em Ciência da Informação no Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília - UNB. E-mail: geozinha@gmail.com

1 - Introdução

A atual conjuntura de mundo globalizado assenta-se em dois pilares básicos: a democracia e a tecnologia. O desenvolvimento desses dois institutos, um no campo da política, outro no campo da ciência, fez com que em praticamente todas as esferas de atividade humana os mecanismos de controle (sejam institucionais, religiosos, artísticos etc.) se tornassem menos reguladores, abandonando a imposição em favor da comunicação. Em relação a esse quadro, Lipovetsky (2004, p. 20) explica que:

Assim opera o processo de personalização, nova maneira de a sociedade organizar-se e orientar-se, nova maneira de gerar os comportamentos, não mais com a tirania dos detalhes, e sim com o mínimo de sujeição e o máximo de escolhas privadas possível, com o mínimo de austeridade e o máximo de desejo possível, com o mínimo de coerção e o máximo de compreensão possível.

Esse novo modelo de comportamento social encontra expoentes também na seara da Ciência da Informação, como, por exemplo, nas iniciativas para assegurar o acesso livre à informação científica. Nesse sentido, em 2002, a UNESCO publicou, através de seu Comitê do Patrimônio Mundial, a Declaração de Budapeste, um documento que defende “[...] a causa do patrimônio mundial pela via da comunicação, da educação, da investigação, da formação e da sensibilização” e lista como um de seus objetivos “desenvolver a comunicação para sensibilizar o público e incentivar a sua participação e o seu apoio ao patrimônio mundial” (UNESCO, 2002, p. 2).

Seguiram-se à Declaração de Budapeste outras iniciativas de definição de princípios e estratégias para se criar um ambiente mais favorável à democratização da informação científica. Em 2003, a Declaração de Berlim dispôs sobre o acesso livre ao conhecimento na ciência e humanidades, focando no reconhecimento do potencial da internet para a representação global e interativa do conhecimento humano e para o acesso irrestrito a conteúdos informacionais; e na defesa de uma Web sustentável, interativa e transparente na qual conteúdos e ferramentas de software fossem livremente acessíveis e compatíveis.

O Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) seguiu a tendência mundial e publicou em 2005 o Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica, documento que apresenta a seguinte constatação:

[...] a comunidade científica enfrenta dificuldades no acesso à informação científica, se considerado o modelo tradicional de publicação científica. Tradicionalmente, os artigos ou trabalhos científicos são publicados em revistas especializadas, e a forma de acesso a esses trabalhos dá-se mediante assinatura das publicações pelas bibliotecas ou pelo pesquisador (IBICT, 2005).

Na literatura de Ciência da Informação é assente o entendimento de que o modelo tradicional de comunicação científica apresenta visíveis sinais de desgastes. Para Moreno e Márdero Arellano (2005, p. 78)

O ciclo do conhecimento científico compreende a produção, a comunicação e a aplicação do conhecimento gerado. Nesse processo, a comunicação exerce uma função essencial, na medida em que viabiliza a divulgação dos resultados de pesquisa e promove a discussão entre os pares. Por um lado, a construção do conhecimento científico depende da disseminação e da transferência das informações, com o intuito de consolidar novos conceitos. Por outro lado, o processo interativo de discussão e aprovação pelos cientistas garante confiabilidade à pesquisa.

Na tentativa de solucionar esse problema e propor uma alternativa a essa situação é que nasceu o movimento Open Archives. Nesse sentido o Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica recomenda à comunidade científica o apoio ao movimento mundial em favor do paradigma do acesso livre à informação científica.

O IBICT segue com uma recomendação imperativa para que as instituições acadêmicas brasileiras criem repositórios institucionais e temáticos e requeiram que seus pesquisadores depositem uma cópia de todos os seus trabalhos publicados em pelo menos um repositório de acesso livre, promovam o apoio à publicação de resultados de pesquisa em periódicos de acesso livre, reconheçam a publicação em ambiente de acesso livre para efeito de avaliação e progressão acadêmica e tenha disponíveis em ambiente livre os periódicos editados pela instituição ou seus órgãos subordinados.

Introduz-se, portanto o conceito de Repositório tanto no nível temático, quanto no nível institucional. Café et al (2003) conceituam repositório temático como um conjunto de trabalhos de pesquisa de uma determinada área do conhecimento, disponibilizados na Internet. Os repositórios utilizam tecnologias abertas e seguem a filosofia da iniciativa dos arquivos abertos, possibilitando maior acessibilidade à produção dos pesquisadores e à discussão entre seus pares. As características principais dos repositórios influem tanto no processo, quanto no produto final da atividade científica, sendo essas características, segundo Café et al (2003): processamento automático dos mecanismos de discussão entre os pares, geração de versões de um mesmo documento, tipologia variada de documentos, auto-arquivamento e interoperabilidade entre todos os repositórios temáticos e seus serviços agregados.

O repositório institucional tem um caráter multidisciplinar, uma vez que agrupará, no âmbito de uma universidade, uma infinidade de tipologias documentais referentes a tantos

quantos forem os temas investigados pelas diversas comunidades e sub-comunidades de pesquisa da instituição.

Para Café et al (2003, p. 4) a função principal do repositório institucional é “[...] preservar e disponibilizar a produção intelectual da instituição representando-a, documentando-a e compartilhando-a em formato digital”. Observa-se que o repositório institucional contribui para a análise quantitativa e qualitativa da produção científica de uma instituição, além de representar uma inovação quanto às possibilidades de tornar o processo de comunicação científica mais transparente, efetivo e democrático.

Os repositórios institucionais são sistemas de recuperação da informação, ao ponto que se enquadram no conceito de Baeza-Yates e Ribeiro Neto (1999) quanto aos sistemas que lidam com representação, armazenagem e organização e acesso a itens informacionais. A literatura é pacífica quanto ao objetivo principal desses sistemas: recuperar documentos relevantes.

O conceito de relevância aplica-se tanto a questões intrínsecas ao usuário, quanto a questões referentes à tecnologia dos sistemas informacionais. Nesse contexto o objetivo deste trabalho é analisar o estado atual do sistema, analisando o conceito de relevância por ele utilizado e propor melhores métodos para a comunicação usuário-sistema no âmbito dos repositórios institucionais. Para isso proceder-se-á uma análise das características da ferramenta de desenvolvimento de repositórios, do conceito de relevância e do conceito de feedback do usuário, no âmbito do Repositório Institucional da Universidade de Brasília.

2 - Repositório Institucional da Universidade de Brasília

O Repositório Institucional é um conjunto de serviços oferecidos pela Biblioteca Central da Universidade de Brasília para a gestão e disseminação da produção científica e acadêmica da comunidade universitária. Todo o seu conteúdo está disponível publicamente e, por estar amplamente acessível, proporciona maior visibilidade e impacto da produção científica da instituição. Sobre os repositórios institucionais, Lynch (2003 apud VIANA; MÁRDERO ARELLANO, 2006, p. 4) salienta que:

[...] um repositório institucional acadêmico é um conjunto de serviços que a universidade oferece para os membros da sua comunidade, para o gerenciamento e disseminação do material digital criado pela instituição e pelos seus membros. É essencialmente o compromisso de uma instituição de cuidar do material digital, incluindo a preservação a longo prazo, quando apropriada, a organização, acesso e distribuição.

O desenvolvimento do Repositório Institucional da Universidade de Brasília - RIUnB faz parte do projeto das Bibliotecas Digitais da UnB - dentre estas a Biblioteca Digital e Sonora, a

Biblioteca Digital de Teses e Dissertações e a Biblioteca Digital de Monografias - e é objetivo estratégico da Universidade, pois permite que toda a comunidade possa acessar a produção intelectual produzida. O RIUnB iniciou-se com uma iniciativa de pesquisa da Profa. Dra. Sely Costa, do então Departamento de Ciência de Informação e Documentação – CID, atualmente Faculdade de Ciência da Informação, da Universidade de Brasília. Para a execução do projeto, a Biblioteca Central contou com a parceria interna da Reitoria, do Decanato de Pesquisa e Pós-Graduação e do próprio Departamento de Ciência da Informação, além do CESPE. Como parceria externa, contou com o apoio do IBICT e da Universidade do Minho de Portugal.

O Repositório Institucional da UnB é constituído de material produzido pelos membros de sua comunidade acadêmica (professores, alunos, entre outros) em termos de artigos de periódicos, livros ou capítulos de livros, trabalhos apresentados em eventos e outros materiais que forem considerados pertinentes e relevantes como produção intelectual da comunidade da UnB. Para garantir a qualidade do material que será disponibilizado é essencial que, antes de se publicar no Repositório, ele tenha sido avaliado pelos pares, no caso de artigos de periódicos e trabalhos apresentados em eventos, afinal “o processo de avaliação de originais pelos pares é, até o momento, o que confere credibilidade ao conhecimento científico divulgado” (STUMPF, 2008, p. 19). Para os demais materiais disponibilizados, a qualidade é garantida com a avaliação crítica daqueles com os mesmos interesses em comum.

Como já explicitado no início, os repositórios institucionais são de extrema importância para a disseminação do conhecimento produzido por uma comunidade científica e, vale destacar as características que o RIUnB e outros Repositórios Institucionais possuem: acesso público transparente, ampla tipologia de documentos, multidisciplinaridade e preservação digital.

2.1 DSpace

O RIUnB, acompanhando a tendência internacional para disponibilizar sua produção acadêmica, optou por utilizar o sistema DSpace. Esta plataforma permite capturar itens em diferentes formatos – texto, vídeo, áudio, e dados – distribuí-los pela web, além de ser possível indexar os documentos para que então os usuários possam pesquisar e recuperar seus itens. Permite, também, a preservação digital dos documentos nele arquivados.

A escolha desse sistema foi baseada no destaque deste sobre as demais ferramentas que podem ser utilizadas como repositórios, pois este tem tido excelente aceitação tanto internacional quanto nacional e também por ser um sistema livre, disponível on-line. O DSpace é um software livre desenvolvido pelo Massachusetts Institute of Technology (MIT) e Laboratórios Hewlett-Packard, é apropriado para a criação e desenvolvimento de repositórios.

Sua arquitetura de software é simples e eficiente, além de ser uma tecnologia de ponta, foi intencionalmente produzido para servir de repositório institucional e direcionado para o acesso aberto à informação científica (VIANA; MÁRDERO ARELLANO; SHINTAKU, 2005).

O DSpace do RIUnB utiliza, para a descrição bibliográfica de seus documentos, o padrão internacional Dublin Core⁵ (DC) em sua versão qualificada, por ser o mais adequado à informação bibliográfica. Esse padrão é de fundamental importância para o compartilhamento de metadados, pois uma das características principais do DSpace é que ele adota o protocolo para coleta de metadados da Iniciativa dos Arquivos Abertos (Open Archives Initiative Protocol for Metadata Harvesting - OAI-PMH).

No Brasil várias instituições utilizam o DSpace, como a Biblioteca Digital Jurídica (BDJur) do Superior Tribunal de Justiça e a Biblioteca Digital do Senado Federal, o que reforça ainda mais a grande aceitação do sistema pela comunidade nacional.

Como parte do processo de recuperação dos documentos, o DSpace oferece a indexação em texto completo que ocorre quando o usuário utiliza a pesquisa rápida ou a busca avançada optando pelo campo “Palavras-chave”. No DSpace a indexação é pré-configurada para pesquisar as primeiras 10.000 palavras, mas pode ser facilmente alterado para fazer a busca em todas as palavras do documento.

Como ponto positivo deste tipo de indexação é possível apontar o fato de que todas as palavras se tornam ponto de recuperação (com exceção das stopwords – palavras não significativas), o que faz com que a recuperação de documentos seja maior. Por outro lado, aumenta a revocação, quantidade de documentos relevantes recuperados dentre o universo de documentos que são relevantes existentes, e diminui a precisão na recuperação dos itens relevantes. Não há, portanto, contextualização na recuperação dos documentos, mas, na falta de opções que possam realizar com maior precisão a busca de documentos, a indexação em texto completo se torna interessante. A saída seria realizar a descrição dos documentos com metadados mais elaborados, descrevendo-os de maneira complexa e completa.

Além da pesquisa, é possível visualizar e recuperar os documentos de diferentes formas: Comunidades e Coleções, Data de publicação, Autor, Orientador, Título e Assunto. Nos resultados da pesquisa rápida ou da pesquisa avançada é possível classificar os documentos de acordo com alguns critérios: relevância, data de submissão, data de publicação e título. O DSpace oferece a ordenação por relevância de acordo com a quantidade de termos buscados que aparecem nos documentos em relação à quantidade de palavras do documento. Será estudado mais à frente que este modelo de relevância por frequência não é o mais

⁵ Para mais informações, acesse: www.dublicore.org.

adequado, pois além de trazer documentos que não são relevantes à pesquisa do usuário, não faz a pesquisa considerando as palavras buscadas como termos e sim como palavras isoladas. Faz-se necessário, portanto, conceituar esse critério de avaliação da informação e identificar aquele conceito que melhor se aplica a recuperação da informação relevante sob a perspectiva do usuário.

3 - O Conceito de Relevância

A informação é um dos produtos de maior valor da sociedade atual. Para Capurro (2007, p. 149), porém, o surgimento da tecnologia da informação e seus impactos globais é que caracterizam a nossa sociedade como uma sociedade da informação. Segundo este autor, “é lugar comum considerar-se a informação como condição básica para o desenvolvimento econômico juntamente com o capital, o trabalho e a matéria-prima, mas o que torna a informação especialmente significativa na atualidade é sua natureza digital.”

O desenvolvimento das tecnologias de informação e comunicação (TIC's) contribui para o crescimento exponencial de informações digitais disponíveis. A facilidade de ferramentas de publicação e consumo de informação estimula a entrada nessa sociedade da informação, pois oferecem a todos, especialistas ou não, aplicações dinâmicas e de baixo custo para disseminar seus trabalhos.

Esse crescimento, entretanto, acontece de forma desestruturada e sem critérios de qualidade, fato que preocupa toda a sociedade. A dificuldade em recuperar informações relevantes, de qualidade e que atendam às reais necessidades do usuário é um dos maiores desafios dos profissionais das diversas áreas que atuam em Sistemas de Recuperação da Informação (SRI), em especial os da Ciência da Informação. Hoje em dia, encontra-se rapidamente muita informação em diversos mecanismos de busca na rede, mas o que realmente é útil precisa ser filtrado muitas vezes, de uma lista bastante extensa, o que onera em demasia todo o processo de recuperação da informação.

A busca por efetividade nos sistemas de recuperação da informação é uma constante e algumas medidas são bastante utilizadas na avaliação desses serviços. A precisão (precision) e a revocação (recall) são duas das medidas mais utilizadas para avaliar o desempenho de um SRI.

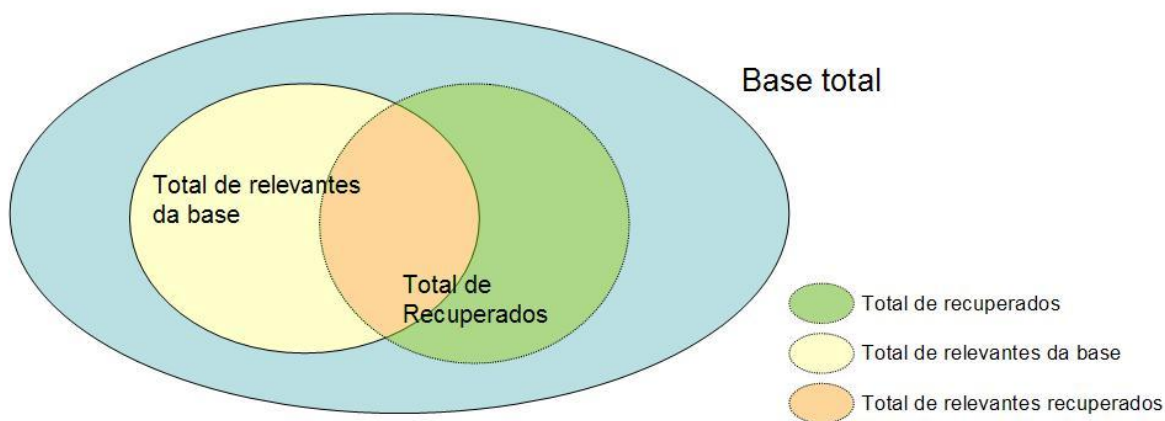


Figura 1: Resultado de busca. Fonte: Elaboração dos Autores.

Conforme Su (1994), essas medidas podem ser definidas de acordo com a seguinte fórmula:

Precisão (*precision*): Nº de objetos informacionais relevantes recuperados pelo sistema

Total de objetos informacionais recuperados

Revocação (*recall*): Nº de objetos informacionais relevantes recuperados pelo sistema

Total de objetos informacionais relevantes da base

O Sistema de Recuperação da Informação ideal é aquele que retorna todos os objetos informacionais relevantes para a requisição feita, ou seja, possui uma revocação igual a 1 e apenas documentos relevantes para a necessidade de informação do usuário, precisão igual a 1.

O Repositório Institucional da UnB em um dos seus serviços oferece uma busca simples por termo livre que recupera todos os documentos que contenham o termo pesquisado em qualquer parte do documento. O resultado pode ser ordenado por ordem crescente e decrescente do título ou por relevância.

Essas medidas de avaliação de SRI e essa forma de ordenação do resultado da busca do repositório remetem à necessidade de conceituar relevância, analisar sua natureza e suas manifestações, para ser possível avaliar a real efetividade do sistema da UnB.

Segundo Figueiredo (1977, p. 75), foi com a publicação de "Sources of information on specific subjects" por S.C. Bradford em 1934, no qual o autor afirmou que "há periódicos de âmbito, obviamente e a priori, relevantes a assuntos investigados", que, pela primeira vez, o conceito de relevância apareceu dentro do contexto de medida de uma fonte de informação.

Desde então, o conceito de relevância vem sendo analisado, mas foi de Cuadra e Katter (apud SARACEVIC, 1975) uma das primeiras definições discutidas na literatura. Segundo os autores, a relevância é uma relação entre uma declaração de informação requisitada por

uma consulta e algum documento contido na base. Nesse caso, relevância seria uma propriedade do sistema e, portanto, depende apenas de como este adquire, representa, organiza e associa as informações (SARACEVIC, 1996).

Para Schamber, Eisenberg e Nilan (1990), a relevância é um julgamento de qualidade entre a informação recuperada e a necessidade de informação do usuário. Nessa definição, o usuário é que determina se o documento recuperado atende as suas necessidades. Relevância aqui refere-se aos contextos subjetivos que são empregados pelo usuário para julgar os objetos informacionais. Aspectos cognitivos, situacionais e psicológicos dos usuários são fundamentais para a compreensão do julgamento.

Os dois conceitos acima definem duas categorias de relevância, citadas na literatura e Schamber (1994). A primeira é a relevância orientada ao sistema e a segunda, relevância orientada ao usuário.

De acordo com Saracevic (2007, p. 1931):

Relevância é como uma árvore de conhecimento. A estrutura básica do sistema de relevância na ciência da informação é uma dualidade: a relevância de tópicos e a relevância do usuário. Cada uma tem seus galhos e devem ser bem exploradas, mas fazem parte da mesma árvore.

A partir dessa classificação e da percepção da necessidade de observar os dois tipos de relevância de forma complementar, Saracevic (1996) propôs um modelo denominado “manifestações de relevância”, no quadro 1, no qual procura detalhar os tipos anteriores.

Manifestações de relevância		
Nível	Relevância	Descrição
Do sistema	Sistêmica ou Algorítmica	Relação entre uma consulta, um processo de busca e os documentos recuperados ou não recuperados.
Do sistema	Tópica	Relação entre o assunto ou tópico expressado e uma consulta e o tópico ou assunto coberto pelos documentos.
Do usuário	Cognitiva/ pertinência	Relação entre o estado de conhecimento do usuário, sua necessidade de informação cognitiva e os documentos recuperados.
Do usuário	Utilidade / aspectos situacionais	Relação entre a situação, tarefa ou problema em mãos (necessidade) e a utilidade do documento percebida pelo usuário.
Do usuário	Motivacional / afetiva	Relação entre as intenções, metas, objetivos e motivações de um usuário e os documentos recuperados.

Quadro 1 – Manifestações de relevância. Fonte: Saracevic (1996).

Ampliando esse conceito de relevância e percebendo suas manifestações, a necessidade de rastreamento e análise de critérios dos usuários para avaliação dos documentos tornou-se urgente.

Pitangueira (2004) consolidou uma série de critérios propostos por diversos autores e classificou segundo o modelo de Saracevic de “manifestações de relevância” com algumas mudanças.

Na manifestação tópica ele propôs uma expansão do conceito de Saracevic e denominou essa manifestação, conforme Borlund (2003), de topicalidade intelectual/assunto. A avaliação da correspondência de tópico/assunto do objeto informacional e o descrito na busca, nesse caso, são feitos pelo usuário e de forma intelectual. Dessa forma, pode-se incluir mais uma medida da percepção do usuário. Alguns dos critérios dessa manifestação são:

Concordância tópica: mede quão correta foi a interpretação dos termos da consulta em relação ao assunto apresentado nos documentos.

Informatividade do abstract: quanto a sinopse foi esclarecedora.

Concordância com a necessidade informacional: mede o quanto as informações do objeto satisfazem com a necessidade do usuário.

Profundidade/escopo: avalia o quanto o documento está focado no assunto de interesse do usuário e quanto o objeto se aprofunda no assunto pesquisado.

Percepção temporal/atualidade: medida para identificar o quão recente e atual é o objeto informacional.

Na manifestação cognitiva/pertinência alguns critérios identificados foram:

Aumento de conhecimento: quanto o conteúdo foi pertinente para acréscimo de conhecimento sobre a necessidade.

Novidade: mede quanto a informação é nova para o usuário.

Ponto de vista: medida que se refere ao quanto as informações reforçam ou não a visão do usuário.

Já na manifestação utilidade/aspectos situacionais, os seguintes critérios são apresentados:

Utilidade: mede quanto o objetivo informacional foi útil para a necessidade de informação.

Resolução de problemas: mede quanto o documento foi útil na resolução de algum problema.

Tomada de decisão: refere-se à utilidade do documento para ajudar a decidir sobre sua necessidade de informação.

Observando os conceitos, tipos, manifestações e critérios de relevância, percebe-se a necessidade dos projetos de SRI contemplarem a participação do usuário. Sendo assim, a criação de mecanismos de interação do usuário com o sistema, emitindo feedback dos objetos informacionais recuperados, é fundamental para a busca de um SRI efetivo. Além disso, o usuário deve participar ativamente da definição do conceito de relevância utilizado pelo sistema.

4 – Feedback do Usuário aos Sistemas de Recuperação da Informação

No contexto da recuperação da informação, o feedback é a resposta que o usuário dá aos sistemas de recuperação da informação. Sua contribuição é de extrema importância, pois auxilia para que a relevância do sistema ao recuperar as informações solicitadas pela questão do usuário (query) se aproxime ao máximo da real necessidade de informação deste usuário (BAEZA-YATES, RIBEIRO-NETO, 1999).

Este feedback, também denominado feedback de relevância é, segundo Croft (1995, p. 2) “o processo onde usuários identificam documentos relevantes em uma lista inicial de documentos recuperados, e o sistema cria uma nova consulta/questão baseada na amostra dos documentos relevantes”.

Várias técnicas para dar o feedback de relevância aos sistemas têm sido pensadas e desenvolvidas principalmente porque o cenário atual é caracterizado por uma grande sobrecarga de informações disponíveis gerando uma grande dificuldade de se localizar informações que realmente sejam relevantes para as necessidades de informação do usuário e que também sejam de qualidade. Muitas vezes o usuário sente-se desestimulado a procurar informações diante da dificuldade de se encontrar o que se procura e, sem dúvida, filtrar toda a informação torna-se essencial.

Para Cazela (2006), os termos “recuperação da informação” e “filtragem da informação” apesar de parecerem bastante semelhantes e ambos procurarem auxiliar na resolução de problemas relacionados à busca de informações, existem diferenças entre eles. A recuperação da informação está relacionada com a busca de documentos textuais que traduzam uma necessidade de informação momentânea do usuário traduzida em uma questão (query). Já a filtragem apresenta uma abordagem distinta que mantém um perfil do usuário baseado em suas preferências. A cada novo item inserido no sistema, checam-se as suas relações com as preferências do usuário, realizando assim, a filtragem.

Várias têm sido as tecnologias desenvolvidas com o intuito de apoiar a seleção, recuperação e filtragem de informações pelo usuário final de sistemas de informação. Diante dessa perspectiva, as recomendações apresentam-se como uma alternativa aos sistemas de recuperação de informação, no intuito de prover uma interação direta entre o usuário e o sistema. As recomendações, baseadas em técnicas de filtragem de informação, “expressam a experiência e opinião de um público, e são o meio pelo qual indivíduos podem se beneficiar da experiência prévia de outros” (GROSSER, 1991 apud MARQUES, 2007).

De acordo com Resnick e Varian (1997), os sistemas de recomendação utilizam as opiniões de uma comunidade de usuários para assessorar indivíduos desta mesma comunidade a identificarem conteúdos de interesse em um conjunto de opções que poderiam caracterizar uma sobrecarga.

Esse tipo de sistema é atualmente bastante utilizado em sites de empresas de comércio eletrônico, como o Amazon.com, com o intuito de aumentar as vendas e atender com mais qualidade o seu usuário.

Segundo Cazela (2006), um dos grandes desafios dos sistemas de recomendação é alcançar casamento correto dos itens que estão sendo avaliados e os usuários que estão recebendo a recomendação em questão, ou seja, definir e descobrir os relacionamentos entre esses interesses.

Várias são as abordagens de recomendação. Elas surgiram da necessidade de se prever os documentos mais relevantes para satisfazer as necessidades de informação do usuário e se apresentam como estratégias para aperfeiçoar os resultados alcançados pelas técnicas de recuperação da informação (MARQUES, 2007). Baseadas nas técnicas de filtragem colaborativa, filtragem de conteúdo e filtragem híbrida, as abordagens serão brevemente caracterizadas a diante.

4.1 - Abordagem de recomendação social ou recomendação colaborativa

Esta abordagem faz o uso da relevância em grupo, ou seja, vários usuários diferentes estabelecem critérios de relevância, que classificam informações de interesse comum (BAEZA-YATES; RIBEIRO-NETO, 1999), e evidencia a troca de experiência entre as pessoas desse grupo. A abordagem de recomendação social parte do princípio de que as melhores recomendações são aquelas realizadas por usuários com gostos semelhantes e utiliza-se da técnica de filtragem colaborativa (collaborative filtering). Nesta técnica, a filtragem de informações é feita com ajuda humana baseada na avaliação de usuários, que dão opiniões e/ou notas para que alguma coisa seja recomendada (MARQUES, 2007).

De acordo com Bergholz (2002 apud CAZELA, 2006) existem duas formas de se extrair as preferências dos usuários nos sistemas de abordagem colaborativa: explícita e implícita. Na forma explícita, o usuário expõe explicitamente as suas preferências; enquanto que na forma implícita, essas preferências são identificadas analisando-se o comportamento do usuário ou quais as ações ele realiza.

A principal diferença entre a filtragem colaborativa e a baseada no conteúdo é que nesta última técnica os itens são filtrados por seu conteúdo, na filtragem colaborativa os itens são filtrados com base na avaliação feita pelos usuários daqueles itens.

4.2 - Abordagens de recomendação baseadas na técnica de filtragem de conteúdo

As recomendações que utilizam a técnica de filtragem baseada no conteúdo (content-based filtering) dão origem a duas abordagens: a abordagem baseada em conhecimento e a abordagem baseada no conteúdo (MARQUES, 2007).

Na abordagem fundamentada em conhecimento são indicados itens ou outros produtos com base em preceitos editados pelo próprio administrador do sistema de recuperação de informações ou também extraídas pela mineração de dados (data mining). No caso do site do Google, por exemplo, ao se buscar por uma palavra-chave qualquer, ao lado irão aparecer vários links patrocinados, determinados pelo sistema, relacionados com o tema buscado. Além disso, poderão aparecer sugestões de se buscar por outras palavras relacionadas (mineração de dados).

Já a abordagem baseada no conteúdo vai realizar recomendações de itens de acordo com o perfil do usuário ou ainda com o histórico deste usuário. Por exemplo, a partir de itens que ele já comprou em um determinado site ou até mesmo pelo preenchimento de um perfil ou lista de interesses quando o usuário se cadastra em algum sistema.

Esse tipo de recomendação é utilizado em sites de busca, como o Vivíssimo, e também freqüentemente em sites de comércio eletrônico. Neste último caso, se assemelham muito aos serviços de disseminação seletiva da informação, onde cada usuário só irá receber informações sobre produtos ou serviços já elencados por ele como sendo de interesse.

A técnica de filtragem baseado no conteúdo, de acordo com Balabanovic e Shoham (1997 apud CAZELA, 2006), objetiva então, gerar automaticamente descrições dos conteúdos de cada item e comparar esta descrição com a descrição das preferências dos usuários ou com o histórico de consumo dos mesmos, visando verificar se o item é ou não relevante para o usuário em questão.

Uma forma de se aplicar esta técnica é solicitar uma análise de itens ao usuário onde ele avaliará os itens sendo de seu interesse ou não. Depois de realizada a avaliação, o sistema buscará por itens cujo conteúdo se assemelha aos interesses do usuário e descartando os que não se assemelham (CAZELA, 2006).

4.3 - Abordagem híbrida

A abordagem híbrida é composta pela combinação da filtragem colaborativa e da filtragem de conteúdo com o intuito de construir um sistema de informações que possa melhor atender os usuários.

Sistemas de informação híbridos, como o site Amazon.com e o Buscapé, diminuem as desvantagens das duas técnicas e unem as vantagens de cada uma para aprimorar a recuperação de informações relevantes para o usuário (MARQUES, 2007). O que pode ser visualizado na figura seguinte.

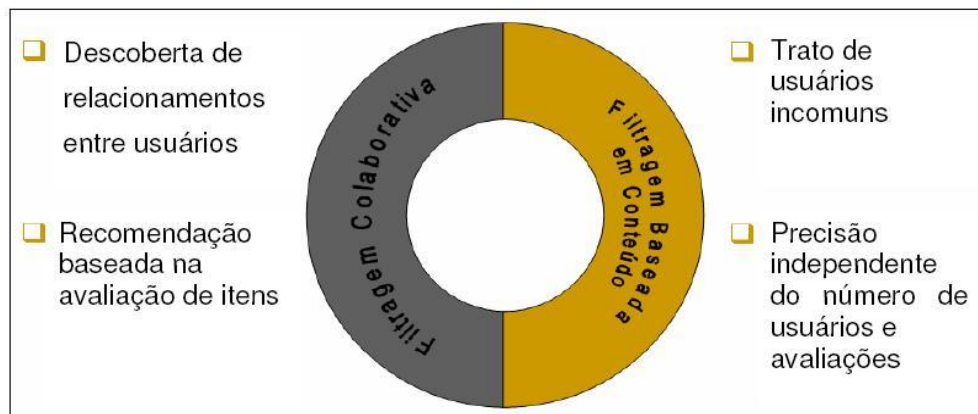


Figura 2: Abordagem híbrida. Fonte: Cazela (2006, p. 35)

5 – Proposta para o RIUnB

Após o estudo feito acima, conclui-se que o critério de relevância adotado pelo DSpace, que é a recuperação da informação pela frequência de palavras, não se constitui no melhor método para atender de maneira mais satisfatória as necessidades reais dos usuários. Partindo do princípio de que o conceito de relevância é muito subjetivo, e que ele depende do julgamento do usuário que demanda uma questão ao sistema, verifica-se que a participação do usuário no processo de recuperação da informação é fundamental.

Sugere-se então, a abordagem de recomendação social ou colaborativa, baseada na técnica da filtragem colaborativa como forma do usuário prover um feedback de relevância ao RIUnb.

Para um melhor aproveitamento dessa abordagem é necessário uma coleta de informações mais detalhadas dos usuários, suas preferências e interesses possibilitando uma boa identificação do perfil dos usuários do repositório. Desta forma, usuários com perfis semelhantes recomendarão itens recuperados pelo sistema utilizando-se da avaliação desses itens, conferindo nota a cada documento considerado relevante.

Associado ao feedback de relevância dado pelo usuário ao sistema, propõe-se, para possibilitar uma recuperação da informação mais relevante para este usuário final, que se configure uma mudança do atual método de busca, baseado na frequência por palavras, para o método vetorial.

Verificou-se que a forma atual do RIUnb atribuir relevância aos documentos não considera o campo específico onde os termos buscados são encontrados. Simplesmente atribui-se ao documento com uma maior freqüência do termo procurado uma relevância maior e ordena-se por esse índice de relevância.

Título	Base de termos	Índice = Frequência do termo buscado na base de termos
Resumo		
Palavras-chave		
Autor		
Texto Completo		

Quadro 2 – Índice de relevância atual do RIUnB.

Fonte: Elaboração dos Autores.

O método vetorial, ou a busca por vetor, possibilita considerar a estrutura de campos disponíveis no documento como partes independentes para a construção do índice de relevância. Cada campo tem um peso ou uma pontuação diferente na elaboração desse índice.

Vetor	Pontuação	
Título	400	Índice = somatório de pontos
Resumo	200	
Palavra-chave	400	
Autor	400	
Texto Completo	50	

Quadro 3 – Índice de relevância proposto para RIUnB.

Fonte: Elaboração dos Autores.

Na distribuição de pontuação acima, o sistema atribuirá importância maior a presença do termo no campo título, palavra-chave e autor do que em resumo e 8(oito) vezes maior que a presença do termo no texto completo do documento.

Além de aproveitar a estrutura de campos do documento, a utilização de vetores possibilita ainda considerar outros fatores para a atribuição da relevância. Propõe-se, então, o acréscimo de dois outros vetores na obtenção do índice. A avaliação positiva dos documentos por parte de usuários de mesmo perfil e a avaliação positiva por parte de usuários de perfil diferente.

Vetor	Pontuação	
Título	400	Índice = Somatório de pontos
Resumo	200	
Palavras-chave	400	
Autor	400	
Texto Completo	50	
Avaliação positiva de usuário com perfil diferente	200	
Avaliação positiva de usuário com mesmo perfil	400	

Quadro 4 – Índice de relevância completo proposto para o RIUnB. Fonte: Elaboração dos Autores.

Considerar essa avaliação separadamente é necessário visto que um objeto informacional pode ser considerado relevante para um usuário por motivos específicos do seu interesse, mas para um usuário com outro interesse o documento pode não ser útil ou relevante.

6 – Conclusão

Este trabalho, de forma alguma, pretende ser exaustivo e profundo o suficiente nas temáticas trabalhadas, pois entendemos que cada uma pode ser desenvolvida a ponto de constituir vários trabalhos independentes. Objetiva, sim, fornecer o entendimento da necessidade de se considerar ambas, a relevância orientada ao sistema e a relevância orientada ao usuário, de acordo com o observado por Schamber (1994), no desenvolvimento de sistemas de informação, como os repositórios institucionais.

Debater sobre repositórios institucionais exige uma compreensão tanto ideológica quanto técnica dos vários aspectos relacionados à questão da comunicação científica. Deve-se ter em mente que esses repositórios são, conceitualmente, Sistemas de Recuperação da Informação imbuídos do espírito do acesso livre e democrático à informação científica.

A discussão sobre os repositórios é relativamente nova e o foco ainda não está na análise desses sistemas sob o prisma da recuperação da informação. A preocupação com o desenvolvimento de melhores modelos teóricos para recuperação da informação é refletida neste trabalho.

Baseados no compartilhamento de informações entre seus usuários, esses sistemas devem considerar a participação do usuário na definição dos critérios de recuperação de

informação. Para isso, propõe-se, para um ambiente novo e ainda em desenvolvimento, como é o Repositório Institucional da Universidade de Brasília, disponibilizado ao público em setembro de 2008, a utilização de uma abordagem de recomendação social, baseada na filtragem de informações de forma colaborativa entre usuários de perfis semelhantes, associada a uma busca baseada em vetores com pesos diferenciados.

Essa abordagem não é a única forma de se fornecer o feedback dos usuários aos sistemas, mas para a proposta em questão se mostrou a mais indicada.

A alteração do conceito de relevância adotado pelo sistema aumenta a possibilidade de se recuperar documentos que realmente satisfaçam as necessidades dos usuários, uma vez que o processo de recuperação no RIUnb passa da busca por frequência de palavras para a busca baseada na frequência por vetores.

Dessa forma, a relevância utilizada pelo sistema passa a ser relevância orientada ao usuário, até então não considerada, e que passa a se utilizar do feedback de relevância fornecida pelo usuário.

Para este usuário, entendido como fundamental no processo de recuperação da informação, faz-se necessário aprimorar as técnicas de feedback adotadas pelo RIUnB, utilizando-se abordagem de recomendação social.

Com a utilização dessa abordagem, usuários com interesse em comum poderão avaliar e recomendar os documentos recuperados, aumentando a possibilidade de a recuperação ser mais efetiva. A troca de informações entre usuários e sistemas de informação possibilita a reformulação da questão inicial pesquisada, tornando-a e ao seu resultado, o mais próximo possível da real necessidade de informação desse usuário.

Dessa forma, o diálogo entre a Ciência da Informação e a Ciência da Computação assume o posto central nos esforços para um melhor entendimento do conceito de relevância dentro dos sistemas e também para o desenvolvimento de estratégias efetivas de comunicação entre o usuário e o sistema, no intuito de aproximar os resultados obtidos na busca com os desejados pelo consulente.

Para o desenvolvimento de pesquisas futuras, sugere-se um mapeamento mais aprofundado do perfil dos usuários cadastrados no sistema, recolhendo informações sobre as suas preferências, entre outras informações. Assim, futuramente, o sistema poderá utilizar-se de uma abordagem de recomendação híbrida, unindo as técnicas de filtragem colaborativa e filtragem de conteúdo.

Sugere-se ainda, estudo detalhado para determinar quais os pesos ideais dos vetores considerados na busca vetorial, lembrando-se que os pesos utilizados neste trabalho foram meramente ilustrativos. O que se sabe, é que os campos (título, resumo, palavras-chaves,

autor e texto completo) onde é feita a busca pelos termos demandados pelas queries dos usuários, devem ser considerados diferentemente.

Além disso, outros vetores podem ser considerados na busca pelo sistema como a quantidade de downloads do documento (e outras informações que o DSpace já fornece aos seus administradores), assim também como outros critérios de relevância do usuário, por exemplo os consolidados por Pitangueira (2004).

Referências

BAEZA-YATES, R; RIBEIRO-NETO, B. Modern information retrieval. Nova Iorque: ACM Press, 1999. cap. 1

BORLUND, P. The Concept of relevance in IR. Journal of American Society for Information Science, v.54, n. 10 , p. 913-925, 2003.

CAFÉ, Lígia et al. Repositórios institucionais: nova estratégia para publicação científica na Rede.In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 26. 2003. Anais... Belo Horizonte: Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, 2003.

CAZELLA, Sílvio César. Aplicando a relevância da opinião de usuários em sistemas de recomendação para pesquisadores. Porto Alegre, 2006. Tese (Doutorado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Universidade Federal do Rio Grande do Sul.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. Perspect. ciênc. inf. [online]. v. 12, n. 1, p. 148-207, 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100012&lng=en&nrm=iso&tlng=pt>. Acesso em: 10 jun. 2009.

CROFT, Bruce. What do people want from information retrieval? D-Lib Magazine, nov. 1995. Disponível em: <<http://www.dlib.org/dlib/november95/11croft.html>>. Acesso em: 05 jun. 2009.

FIGUEIREDO, Laura Maia de. O conceito de relevância e suas implicações. Ciência da Informação, Rio de Janeiro, v. 6, n. 2, p. 75-78, 1977.

LIPOVETSKY, Gilles. Os tempos hipermodernos. Tradução Mário Vilela. São Paulo: Ed. Barcarolla, 2004

MARQUES, Tiago Miranda. Abordagens de recomendação para a recuperação de perfis: uma proposta de modelo. Brasília, 2007. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília.

MORENO, Fernanda Passini; MÁRDERO ARELLANO, Miguel Ángel. Artigo publicação científica em arquivos de acesso aberto. Arquivística.net. Rio de Janeiro, v. 1, n. 1, p. 76-86, jan./jun. 2005. Disponível em: <<http://www.arquivistica.net>>.

PITANGUEIRA, Antônio Maurício da Silva. Caracterização dos fatores de relevância no processo de recuperação de informação. Porto Alegre, 2004. Dissertação (Mestrado em Administração) – Programa de Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*. New York, v. 40, n. 3, p. 55-58, mar. 1997.

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 the American Society for Information Science and Technology*, v. 58, n. 3, p. 1915-1933, 2007.

SARACEVIC, T. Relevance Reconsidered. In: *Information science: Integration in Perspectives. Proceedings of The Second Conference on Conceptions of Library and Information Science (COLIS 2)*, p. 201-208, October, 1996.

SARACEVIC, T. Relevance: a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, v. 39, n. 4, p. 235-351, 1975.

SCHAMBER, L.; EISENBERG, M.; NILAN, M. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing and Management*, v. 26, n. 6, p. 755-776, 1990.

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

STUMPF, Ida. Avaliação pelos pares nas revistas de comunicação: visão dos editores, autores e avaliadores. *Perspectivas em Ciência da Informação*, v. 13, n. 1, p. 18-32, jan./abr. 2008.

SU, L.T. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, v. 45, n. 3, p. 207-217, 1994.

VIANA, C. L. M.; MÁRDERO ARELLANO, M. A; SHINTAKU, M. Repositórios institucionais em ciência e tecnologia: uma experiência de customização do DSpace In: *SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS*, 3., 2005, São Paulo. Anais ... São Paulo: CRUESP, 2005. Disponível em: <<http://bibliotecas-cruesp.usp.br/3sibd/docs/viana358.pdf>>. Acesso em: 10 jun. 2009.

VIANA, Cassandra Lúcia de Maya; MÁRDERO ARELLANO, Miguel Angel. Repositórios institucionais baseados em DSpace e EPrints e sua viabilidade nas instituições acadêmico-científicas. In: *SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS*, 14., 2006, Salvador. Anais... Salvador: UFBA, 2006.

UNESCO. Comitê do Patrimônio Cultural. Declaração de Budapeste sobre o Patrimônio Cultural. Budapeste: UNESCO, 2002. Disponível em: <www.unesco.pt/pdfs/docs/DecBudapeste.doc>. Acesso em: 08 jun. 2009.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. Manifesto brasileiro de apoio ao acesso livre à informação científica, 2005. Disponível em: <kuramoto.files.wordpress.com/.../manifesto-sobre-o-acesso-livre-a-informacao-cientifica.pdf>. Acesso em: 08 jun. 2009.