# Colaboração na Web como ferramenta para a Recuperação da Informação

**Fábio Augusto Guimarães Teixeira[1], Cláudio Gottschalg-Duque[2]**

[1]Mestrando em Ciência da Informação – Programa de Pós-Graduação em Ciência da Informação – Universidade de Brasília (UnB)

[2]Professor Adjunto (Orientador da pesquisa) – Programa de Pós-Graduação em Ciência da Informação – Universidade de Brasília (UnB)

{fagt,klauss}@unb.br

***Ano de ingresso no programa:*** *2008*

***Época esperada de conclusão:*** *Dez/2009*

***Etapas já concluídas:***

- Seminário de andamento: Disciplina Pesquisa Orientada em Ciência da Informação concluída no 1º semestre de 2008
- Defesa da proposta: Trabalho apresentado/publicado em congresso: Collaborative Indexing on the Web. In: IADIS International Conference WWW/Internet 2008, 2008, Freiburg, Alemanha. Proceedings...

***Abstract.*** *The establishment of Internet and the Web environment has permitted information sharing in a way that has never been seen before. But it has also deepened the problem of finding useful information in information retrieval systems. This research aims at conceiving an information retrieval service that uses the collaborative opportunities of the Web environment to allow users to share representations about information and collaborate so that these can supply access points in future retrieval attempts.*

***Keywords:*** *Information retrieval, Web collaboration, information representation*

***Resumo.*** *O advento da Internet e do ambiente Web possibilitou o compartilhamento de informação de uma forma jamais vista anteriormente. Porém, também intensificou o problema de encontrar informação útil em sistemas de recuperação da informação. Esta pesquisa busca a concepção de um serviço de recuperação da informação que aproveite as possibilidades colaborativas do ambiente Web para permitir que usuários compartilhem representações de informação e colaborem entre si para que essas possam servir de pontos de acesso em tentativas posteriores de recuperação.*

***Palavras-chave:*** *Recuperação da informação, colaboração na Web, representação da informação*

## 1. Fundamentação téorica

O desenvolvimento constante de tecnologias digitais de comunicação permite o crescente acesso à informação. O maior fruto deste desenvolvimento é, provavelmente, o advento da Internet e, posteriormente, da World Wide Web (também conhecida como WWW, ou simplesmente Web). A Web foi concebida por Tim Berners-Lee como um espaço informacional onde as pessoas podem compartilhar informações e colaborar no sentido de criar novo conhecimento. Seria também um ambiente em que todos, sejam pessoas ou máquinas, podem ter acesso a tudo, através da compreensão do conteúdo das informações [Berners-Lee 2000].

Segundo Baeza-Yates e Ribeiro-Neto (1999) "a Web está se tornando um repositório universal de conhecimento e cultura humana, que tem permitido compartilhamento de idéias e informação sem precedentes". Porém, a Web introduziu o problema de encontrar informação útil, o que fez aumentar o interesse em pesquisas em recuperação da informação.

O termo recuperação da informação foi cunhado por Mooers (1951) designando a área responsável pelos aspectos intelectuais da descrição de informação e especificidades para a busca, incluindo sistemas, técnicas ou máquinas empregadas para este fim. A recuperação da informação é encarregada da representação, armazenagem, organização e acesso a itens informativos [Baeza-Yates e Ribeiro-Neto 1999].

De acordo com sua definição, estudos em recuperação da informação podem envolver aspectos distintos como comportamento e necessidades informacionais do usuário, interface, estruturas de representação e armazenamento da informação, algoritmos de recuperação, arquitetura de sistemas, entre outros. Por isso, a área se caracteriza como sendo de pesquisa interdisciplinar.

As tecnologias que compõem a Internet e a Web foram desenvolvidas sob estudos da área de Computação. Com isso, nada mais natural que, a partir de notada a necessidade de melhorias em recuperação da informação neste ambiente, pesquisadores da própria área buscassem novas soluções para este fim. Entretanto, progressos em algoritmos, estruturas de armazenamento e sistemas resultam em aumento de desempenho, mas não necessariamente em melhor recuperação da informação.

Existe um componente fundamental que precisa ser considerado ao se buscar resolver problemas de recuperação da informação e que não se encontra no domínio de máquinas e sistemas. Este elemento é o usuário, responsável não apenas por formular representações de informação (no papel de fonte) ou de busca (no papel de interessado em outros registros), mas também por designar o caráter subjetivo de relevância aos resultados apresentados em uma busca. Como é o próprio ser humano que define qualidade para resultados de consultas, nada mais natural que características cognitivas como comportamento e contextualização sejam alvo de modelagem específica para recuperação da informação [Rijsbergen 1979; Duque 2005; Rosenfeld e Morville 2006].

Exemplificando o caráter interdisciplinar da recuperação da informação, estudos de usuário são desenvolvidos por outras áreas, como a Ciência da Informação [Hjørland 1998]. Esta disciplina emprega conceitos de lógica, lingüística, psicologia, comunicação, biblioteconomia (de onde é derivada) para enfrentar os desafios de estudo

da informação [Borko 1968]. No caso específico da recuperação da informação, também a Ciência da Informação se ateve por bastante tempo às abordagens de melhoria física das aplicações do campo, com pouco foco no usuário. Porém, é possível extrair de seus vínculos com a lingüística, com a comunicação e com a biblioteconomia alguns caminhos para o desafio de encontrar informação útil na Internet, destacando-se a experiência em representação da informação e os estudos do emprego da linguagem natural, de difícil concepção para máquinas [Lancaster 2004; Moens 2000].

Voltando às características do ambiente Web, é necessário ressaltar a possibilidade de colaboração entre seus usuários no compartilhamento de informação e conhecimento. Aplicações colaborativas se tornam cada vez mais comuns na Web. Ao conjunto destes serviços se convencionou chamar de Web 2.0, termo criado pela O'Reilly Media [O'Reilly 2005]. Segundo o autor, essas aplicações podem ser caracterizadas como "software que se torna melhor quanto mais pessoas o utilizam", onde o "comportamento do usuário não é predeterminado", com "rica experiência do usuário" e "confiança nos usuários", entre outros fatores. Um dos maiores exemplos e que encerra todas estas características é a Wikipedia, enciclopédia cujos verbetes são descritos (incluídos, corrigidos ou ampliados) pelos próprios usuários.

## 2. Caracterização da contribuição

De forma sintética, a pesquisa aborda o problema da recuperação de informação útil na Web, determinado pela grande quantidade de registros disponíveis no ambiente e pelos métodos existentes de recuperação, baseados em princípios de indexação e algoritmos de busca que não conseguem representar corretamente as necessidades informacionais dos usuários que buscam informação.

A proposta de trabalho é a concepção de um serviço de recuperação da informação em ambiente Web que permita que usuários compartilhem descrições a respeito de itens recuperados. De forma similar aos verbetes da Wikipedia, cada usuário seria estimulado a descrever textualmente da forma que desejar qualquer item que sua pesquisa tenha retornado. Estas descrições podem partir desde etiquetas curtas (como *tags*) até textos completos em linguagem natural.

Do ponto de vista da Ciência da Informação, tal serviço se caracterizaria pela expansão da representação da informação aplicada sobre cada item. Isto é, ao invés de apenas oferecer a representação primária de um objeto (no caso, o próprio objeto) ou a representação secundária gerada por algum processo de indexação (automático ou manual), o sistema possibilitaria que diferentes usuários registrem suas impressões a respeito do objeto. Tais descrições não apenas ficariam disponíveis para que outros usuários as utilizassem como ponto de acesso em pesquisas posteriores, mas também as alterassem conforme desejado. Trata-se, portanto, de tornar disponível novo conhecimento a respeito de itens informativos.

Estas impressões podem ser coincidentes ou não. Caso não sejam, diferentes sentidos podem ser associados ao mesmo objeto. Da mesma forma como grupos distintos de usuários podem ter interpretações e conceitos diferentes de relevância a respeito do mesmo documento.

Uma vez caracterizada como um serviço Web 2.0 típico, a abordagem se

diferencia dos serviços prestados por sistemas de recomendação ou de filtragem da informação. Enquanto estes são apoiados em algum algoritmo computacional aplicado sobre a armazenagem, manipulação e/ou apresentação de registros, a abordagem colaborativa utiliza exclusivamente contribuições de usuários para estes fins. Por isso, dispõe de recursos numerosos, capazes de registrar conhecimento em um grande volume em curto período de tempo, inclusive possibilitando acompanhar a evolução do conhecimento a respeito daquela informação.

Como a hipótese da contribuição é de que a colaboração entre os usuários do serviço pode facilitar a recuperação de itens informativos úteis na Web, pretende-se efetivamente lançá-lo naquele ambiente como um serviço voltado para um determinado público alvo (por exemplo, interessados no compartilhamento de *links* de serviços Web 2.0). A aplicação permitiria que os usuários buscassem, inserissem, editassem ou excluíssem não apenas os *links* em questão, mas também descrições do que apontam.

Dos diferentes componentes necessários para a definição da arquitetura do serviço de informação proposto, maiores cuidados estão sendo dispensados quanto à escolha da base de dados (registros indexados para recuperação), à forma de gravação dos registros secundários (metadados), à interface de usuário e à política de colaboração, pois esta comporá a estrutura básica quando de seu lançamento. Todo o conteúdo restante deverá ser provido por seus usuários. Como estratégia de lançamento se prevê a divulgação do serviço através do uso de redes sociais na própria Web, de forma a utilizar o próprio meio em que o projeto foi idealizado.

## 3. Estado atual do trabalho

O projeto se encontra em fase de validação e refinamento do modelo a ser empregado pelo sistema de recuperação da informação. Após obter o embasamento teórico que fundamenta não apenas a contribuição da proposta mas também as contribuições de diferentes áreas sobre o assunto e definir a forma como o serviço de recuperação da informação proposto será concebido, é necessário ratificar se o método proposto engloba o conjunto de fatores (sujeitos e materiais envolvidos, tarefas, roteiro, contexto, variáveis controladas e não controladas) necessários para a análise do contraste entre a abordagem proposta pelo modelo e as abordagens estabelecidas atualmente.

## 4. Trabalhos relacionados

É necessário destacar dois trabalhos relacionados ao projeto atual. O primeiro serviu de inspiração para o modelo. O segundo é um serviço de busca na Web lançado por terceiros em 2008 (após o início desta pesquisa) que implementa uma significativa parte dos conceitos abordados.

### 4.1. *User-assigned terminology* [Besser 1997]

Ao analisar o custo para designar termos para cada imagem nas coleções de um banco de dados de imagens, o autor propôs a terminologia designada por usuário (no original, *user-assigned terminology*) da seguinte forma:

> Se nós pudermos desenvolver sistemas para terminologia designada por usuário, gerentes de coleções podem contar com usuários para designar termos ou palavras chave para imagens individuais. Sob tal sistema, quando usuários

> encontram uma imagem, o sistema perguntaria a eles quais palavras eles podem ter usado para buscar essa imagem. Essas palavras são então inseridas no sistema de recuperação e usuários subseqüentes buscando essas palavras irão encontrar a imagem. À medida que o número de pessoas usando tal sistema aumenta, também aumenta o número de pontos de acesso para muitas dessas imagens.

Embora não se tenha mais notícia sobre o desenvolvimento da proposta do autor, a abordagem de solicitar aos usuários alguma informação sobre a busca e depois disponibilizá-la para buscas de outros usuários pode ser considerada como um primórdio do princípio de colaboração, que baseia as aplicações Web 2.0 e este trabalho.

### 4.2. Wikia Search

O serviço de busca Wikia Search, de propriedade da empresa Wikia, Inc. aplica de forma direta conceitos colaborativos da plataforma Wiki para que os usuários realizem buscas e possam incluir sugestões de resultados, sugestões de outros termos relacionados à busca, ou mesmo editar as descrições dos resultados apresentados. Esses usuários podem editar, complementar ou eliminar a descrição realizada inicialmente, opinar sobre a relevância do item para a pesquisa, fazer anotações de uso pessoal ou enviar comentário aos demais usuários que contribuíram para aquela representação.

Embora recente, é possível observar algumas ressalvas ao uso do serviço. A maior delas refere-se à base de dados empregada, que não prevê distinção entre a massa descrita pelos usuários e a de indexação tradicional proveniente dos próprios documentos. Em outras palavras, o usuário, ao realizar uma busca, não sabe se as respostas são provenientes de contribuições de outros usuários. Além disso, a considerar pelo nível das respostas atualmente, ainda são poucas as contribuições de usuários, o que é fundamental para a efetividade do conceito. Mesmo assim, o serviço se configura como bom exemplo inicial da abordagem proposta pelo projeto.

## 5. Avaliação dos resultados

Por se tratar de uma abordagem cujo enfoque é a representação colaborativa de usuários para facilitar a recuperação de informação, se faz imprescindível a participação de um grupo de usuários na etapa de avaliação do projeto. É possível destacar fatores que deverão receber maior atenção quando da ponderação dos resultados e que por isso representam os desafios do trabalho: facilidade de uso e estímulo para colaboração, relevância e potencial da proposta e se permite melhorar a recuperação de forma efetiva.

Considerando estes fatores, algumas dificuldades de implementação e execução do projeto podem ser descritas: construção de interface que facilite o envolvimento dos usuários e permita liberar seus potenciais colaborativos, escolha de um segmento (ou grupo) inicial de usuários para aplicação do conceito, e, principalmente, a seleção de métricas que permitam demonstrar o grau de utilização e satisfação dos usuários do serviço, mesmo considerando que entre elas constarão variáveis extremamente subjetivas. Além disso, outros fatores considerados são a forma de armazenamento dos registros criados e as próprias diretivas deste armazenamento (qual tamanho previsto, em que lugar, por exemplo).

Até o momento, as métricas previstas para avaliação do serviço incluem a

quantidade de usuários registrados, a quantidade de usuários registrados ativos (que tenham efetivamente realizado alguma contribuição), a quantidade de acessos (de usuários registrados ou não), a quantidade de registros inseridos, a quantidade total de colaborações realizadas aos registros inseridos, a quantidade de buscas realizadas e a quantidade de *links* externos acessados a partir dos registros inseridos. Entretanto, esta análise quantitativa deverá ser seguida por outra qualitativa realizada junto aos usuários cadastrados, envolvendo o questionamento da opinião a respeito da relevância do serviço, de seu potencial de crescimento e de pontos positivos e negativos da proposta. A partir desta análise, será possível definir a pertinência do projeto (mantendo-o ativo ou não) e/ou possíveis melhorias visando a sua evolução.

## Referências


Baeza-Yates, R. e Ribeiro-Neto, B (1999), Modern information retrieval, ACM Press.

Berners-Lee, T (2000), Weaving the Web, HarperCollins Publishers Inc.

Besser, H. (1997), "Image Databases: The First Decade, the Present, and the Future", In: *Digital Image Access & Retrieval* (Papers presented at the 1996 Clinic on Library Applications of Data Processing, March 24-26, 1996), Heydorn, P. e Sandore B. (eds.), pp. 11-28. University of Illinois.

Borko, H. (1968) "Information Science: What is it?" In *American Documentation*, v. 19, p. 3-5.

Duque, C. (2005), "SiRILiCO: Uma Proposta para um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia", Dissertação (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, 118 p.

Hjørland, B (1998), "Information Retrieval, Text Composition, and Semantics" In *Knowledge Organization*, v. 25, p. 16-31. Disponível em: <http://dlist.sir.arizona.edu/445/01/ir_semant_2.pdf>. Acesso em 20 nov. 2008.

Lancaster, F (2004), Indexação e Resumos: Teoria e prática, Briquet de Lemos, 2a. edição

Moens, M.-F. (2000), Automatic indexing and abstracting of document texts, Kluwer Academic Publishers.

Mooers, C. (1951) "Zatocoding applied to mechanical organization of knowledge" In *American Documentation*, v. 2, p. 20-32.

O'Reilly, T. (2005) "What is Web 2.0. Design patterns and business models for the next generation of software". O'Reilly Network. Disponível em <http://www.oreillynet.com/lpt/a/6228>. Acesso em 2 dez. 2008

Rijsbergen, C. (1979), Information Retrieval, Butterworths, 2a. edição. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em 2 dez. 2008.

Rosenfeld, L. e Morville, P. (2006), Information Architecture for the World Wide Web, O'Reilly & Associates, Inc., 3a. edição.

Wikia, Inc. (2008), Wikia Search. Disponível em: <http://www.wikiasearch.com>. Acesso em 2 dez. 2008.