DOI: 10.5433/1981-8920.2013v18n3p27

MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO E WEB SEMÂNTICA: A QUESTÃO DA RELEVÂNCIA

LOS MODELOS DE RECUPERACIÓN DE LA INFORMACIÓN Y LA WEB SEMÁNTICA: LA CUESTIÓN DE LA PERTINÊNCIA

Renata Eleutério da Silva - renata_biblio@yahoo.com.br
Mestranda do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP/Marília).

Plácida Leopoldina Ventura Amorim da Costa Santos - placidasantos@gmail.com
Livre-docente em Catalogação pela Universidade Estadual Paulista (UNESP/Marília). Docente do Departamento de Ciência da Informação da UNESP/Marília.

Edberto Ferneda - edbertof@terra.com.br
Pós-doutor em Sistemas de Informação pela Universidade Federal da Paraíba (UFP). Docente do Departamento de Ciência da Informação da UNESP/Marília.

RESUMO

Introdução: A preocupação com a recuperação de informações em sistemas computacionais precede o desenvolvimento dos primeiros computadores pessoais. Modelos de recuperação de informações foram e são até hoje muito utilizados em bases de dados específicas de um domínio, cujo escopo é conhecido. No ambiente Web, existe a necessidade de maiores cuidados no que diz respeito ao tratamento descritivo e temático das informações.
Objetivos: Verificar como a questão da relevância é tratada nos principais modelos computacionais de recuperação de informação e, sobretudo, como o tema é abordado em relação ao futuro da Web, a chamada Web Semântica.
Metodologia: Pesquisa bibliográfica.
Resultados: Nos modelos clássicos estudados neste artigo, percebeu-se que a

preocupação principal é a recuperação de documentos cuja descrição esteja mais próxima da expressão de busca utilizada pelo usuário, o que não necessariamente implica no que este realmente necessita. Na recuperação semântica há o uso de ontologias, recurso que estende a busca do usuário para uma gama maior de possíveis opções relevantes.
Conclusões: A relevância, sendo algo subjetivo e inerente ao julgamento do usuário, dependerá da interação do mesmo com o sistema e, principalmente, ao que de fato ele espera recuperar em sua busca. Os sistemas que se baseiam em um modelo de relevância não são populares, por exigir maior interação e depender da disposição do usuário. A Web Semântica é, até então, a iniciativa mais eficiente no que tange a recuperação de informação no ambiente digital.

Palavras-chave: Informação e Tecnologia. Modelos de Recuperação de Informação. Web Semântica. Relevância.

1 INTRODUÇÃO

A recuperação de informação é um campo da Ciência da Computação que se preocupa em desenvolver e estudar os aspectos relativos à eficiência e eficácia das buscas em um sistema computacional, de modo que os resultados de tais buscas sejam relevantes ao usuário do sistema e, sobretudo, coerentes com sua expressão de busca. Essa temática se faz muito importante à Ciência da Informação, devido aos aspectos ligados ao tratamento e representação das informações, estejam elas em ambiente digital ou não.

Sant'Ana (2008, p. 145) afirma que "com a adoção maciça das tecnologias de informação e comunicação, o volume de informações armazenadas e disponíveis para acesso vem crescendo de forma exponencial" e segue afirmando que, para que essa grande quantia de informações seja transmitida ao usuário da melhor forma, são necessários processos de recuperação cada vez mais eficientes. Desta forma, o aumento nos fluxos informacionais, gerados pela evolução da internet, torna fundamental o desenvolvimento e a melhoria constante de mecanismos de busca e recuperação nesses ambientes informacionais.

A Web se configura hoje como a maior base de dados existente e, consequentemente, a que necessita de maiores cuidados no tratamento descritivo de suas informações. Para que as informações contidas na Web possam ser recuperadas por motores de busca (*search engines*), documentos e páginas devem ser devidamente indexados, processo que é realizado automaticamente por agentes de *software* denominados "*Crawlers*" (rastreadores).

Para a modelagem de bases de dados específicas e locais, alguns modelos clássicos de recuperação de informação são facilmente implementados. Entretanto, tais modelos nem sempre podem ser adaptados para o ambiente Web. Com base na enorme quantidade de informações existentes na Web atual, buscou-se verificar como a questão da relevância é tratada nos principais modelos de recuperação de informação e, sobretudo, como o tema é abordado em relação ao futuro da Web, fase denominada "Web Semântica".

2 MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO

Os sistemas de recuperação de informação, de acordo com Ferneda (2012, p. 13), têm por função "representar o conteúdo dos documentos do *corpus* e apresenta-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação [...]." Os modelos de recuperação de informação vêm sendo desenvolvidos há muitas décadas, muito antes até da invenção do computador e dos recursos tecnológicos que se tem atualmente. Mesmo tendo sido pensadas há muito tempo, muitas ideias relacionadas à recuperação de informação são até hoje utilizadas como base no desenvolvimento de novos sistemas computacionais, que buscam ser cada vez mais eficientes e preocupados com a qualidade das informações que serão recuperadas.

São considerados modelos clássicos de recuperação de informação: o Modelo Booleano, o Modelo Vetorial e o Modelo Probabilístico. Este tópico se destina a abordar esses modelos, explanando de forma sucinta suas principais características. Além desses, serão comentados modelos derivados de pesquisas mais recentes na área da computação, tais como o modelo Fuzzy, as Redes Neurais e os Algoritmos Genéticos.

2.1 Modelo Booleano

O Modelo Booleano é um dos mais utilizados nos sistemas de recuperação de informação, por sua simplicidade e formalismo claro, o que o torna mais facilmente implementável. É baseado na teoria dos conjuntos e na Álgebra de Boole (KURAMOTO, 2002). Em um sistema booleano, o conteúdo informacional dos

documentos é representado por um conjunto de termos de indexação. As buscas são formuladas por meio de uma expressão booleana composta por termos ligados através dos operadores lógicos (AND, OR e NOT). O resultado de uma busca é composto por um conjunto de documentos cuja representação satisfaz às restrições lógicas da expressão de busca.

De acordo com Souza (2006, p. 166), a principal desvantagem do modelo booleano está ligada ao fato de o modelo trabalhar o conceito de relevância de forma binária, ou seja, "os documentos são analisados sob o critério dualista relevante/não relevante, e não é criada nenhuma espécie de ordenação dos resultados que atendam às condições de consulta." Isto é, o resultado de uma busca booleana se caracteriza por uma simples partição *corpus* documental em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem. Presume-se que todos os documentos recuperados são de igual utilidade para o usuário, não havendo mecanismo algum pelo qual os documentos possam ser ordenados (FERNEDA, 2012).

Apesar das desvantagens apontadas, é um modelo com ampla utilização em bancos de dados relacionais e catálogos de bibliotecas, por exemplo.

2.2 Modelo Vetorial

Também conhecido como Modelo Espaço Vetorial, o modelo Vetorial "baseia-se na comparação parcial entre a representação dos documentos e a da consulta do usuário" (KURAMOTO, 2002), o que somente é possível devido à atribuição de peso tanto aos termos da expressão de busca como aos termos de indexação que representam os documentos. Um documento é representado por um conjunto de termos de indexação, cada qual associado a um valor numérico entre 0 e 1, que representa a relevância do respectivo termo na representação do conteúdo informacional do documento. Uma expressão de busca é também representada por um conjunto de termos e seus respectivos pesos, que representam a importância do termo na expressão de busca.

A homogeneidade na forma das representações tanto dos documentos como das expressões de busca permite criar um sistema no qual é possível não só calcular o "grau de similaridade" entre uma expressão de busca e cada um dos

documentos do *corpus,* mas também verificar a semelhança entre dois documentos. O resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade entre a expressão busca do usuário e cada um dos documentos do *corpus*.

Diferentemente da dualidade do modelo booleano, no modelo vetorial o conceito de relevância é tratado como um *continuum* representado numericamente por meio de um número real entre zero e um. Esta característica permitiu o desenvolvimento de diversas técnicas de recuperação de informação utilizadas até hoje, tais como *clustering* (agrupamento), *relevance feedback*, classificação, reformulação da expressão de busca etc., que foram materializadas no sistema SMART (*System for the Manipulation and Retrieval of Text*), desenvolvido por Salton nos anos 60 (FERNEDA, 2012).

2.3 Modelo Probabilístico

Como o próprio nome sugere, o modelo Probabilístico se pauta na teoria matemática das probabilidades. Como afirma Souza (2006), esse modelo supõe que exista um conjunto ideal de documentos que atende a cada uma das possíveis buscas que podem ser feitas no sistema. A partir do primeiro conjunto de documentos resultantes de uma busca, o usuário seleciona alguns que considera relevantes para responder à sua necessidade de informação. A expressão de busca, juntamente com os documentos que foram selecionados como relevantes, é submetida novamente ao sistema de informação, procurando refinar a busca e tentando aproximar-se cada vez mais do conjunto ideal de documentos. Este processo interativo é conhecido como *Relevance Feedback*.

Segundo Ferneda (2012, p. 52):

> O processo de recuperação de informação é caracterizado por seu grau de incerteza no julgamento de relevância dos documentos em relação à expressão de busca. Assim sendo, é mais realístico pensar em uma probabilidade de relevância do que em uma pretensa relevância exata, como a utilizada nos modelos booleano e vetorial.

Segundo o mesmo autor, outra virtude do modelo probabilístico está em reconhecer que a atribuição de relevância é uma tarefa do usuário. É o único modelo

que incorpora explicitamente o processo de *relevance feedback* como base para a sua operacionalização.

2.4 Modelo *Fuzzy*

Este modelo está baseado na lógica *fuzzy*, cujo objetivo é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa (SHAW; SIMÕES, 1999). O modelo *Fuzzy* é baseado no mundo real, no qual a imprecisão e a incerteza são intrínsecas à recuperação de informações (PERES; BOSCARIOLI, 2002). Este modelo visa a superar as limitações do modelo booleano, porém sua discussão é mais restrita à literatura que se dedica à teoria *fuzzy*, não sendo comumente implementada na área da recuperação de informação.

Souza (2006, p. 166) afirma que no modelo em questão:

> [...] busca-se estender o conceito da representação dos documentos por palavras-chave, assumindo que cada *query* determina um conjunto difuso e que cada documento possui um grau de pertencimento a esse conjunto, usualmente menor do que 1. O grau de pertencimento pode ser determinado pela ocorrência de palavras expressas na *query*, tal como no modelo booleano, mas pode também utilizar um instrumento – como um tesauro – para determinar que termos relacionados semanticamente aos termos índice também confiram algum grau de pertencimento ao conjunto difuso determinado pela *query*.

Entendendo a "*query*" como a expressão de busca pela qual o usuário realiza suas pesquisas nos sistemas de recuperação de informação, a lógica *Fuzzy* leva em conta, então, não os resultados inteiros, mas sim os valores dos intervalos entre 0 e 1.

2.5 Redes Neurais

As redes neurais artificiais (RNA) visam a simular computacionalmente o funcionamento biológico dos neurônios cerebrais. Os modelos baseados em redes neurais se propõem simular a atuação do sistema nervoso humano em sistemas de recuperação de informações.

De acordo com Ferneda (2006), em um sistema de recuperação de informação, de um lado estão as expressões de busca do usuário, do outro os documentos e no centro os termos de indexação. Tal composição pode ser vista como uma rede neural, pela semelhança existente entre as estruturas.

Na recuperação de informação, para Souza (2006), as redes neurais se utilizam de padrões para relacionar as expressões de busca dos usuários com os documentos de um acervo, de modo que cada expressão de busca libera um sinal que ativa os termos do sistema e que se propaga aos documentos relacionados. Tais estímulos retornam os sinais a novos termos, em interações sucessivas. As respostas apresentadas ao usuário são definidas por meio desse processo, que podem conter até termos que não foram utilizados na busca, mas que demonstraram ter relação com a expressão pesquisada.

Em um processo de recuperação de informação que utilize as RNA, o resultado final de uma busca será o grupo de documentos do *corpus* que foram ativados no processo de inferência do sistema. De acordo com Ferneda (2012, p. 94), o resultado será composto por documentos diretamente ligados à expressão de busca do usuário e também por documentos que o sistema inferiu durante a coleta pelo fato de possuir algum grau de relevância em relação à necessidade do usuário. Essa característica é uma das vantagens do modelo, já que pode produzir resultados inesperados, que não possuem nenhum termo em comum com as expressões utilizadas, mas que também podem satisfazer à busca do usuário.

2.6 Algoritmos Genéticos

Assim como as Redes Neurais, os Algoritmos Genéticos possuem suas bases na biologia, mais especificamente na genética, partindo da tentativa de representar matematicamente a teoria da evolução das espécies. Pesquisas foram desenvolvidas para que se pudessem adaptar os fenômenos da natureza a modelos a serem utilizados em sistemas computacionais. Partindo desse pressuposto, os Algoritmos Genéticos são técnicas utilizadas para simular o processo de evolução natural, de modo a gerar soluções a um determinado problema. Cada vez que o algoritmo se repete em um processo, são criadas novas estruturas por meio da troca de informações, de modo que as próximas “gerações” sejam cada vez mais aptas a

resolver os problemas de uma dada situação. A apresentação dos resultados mais relevantes dependerá da interação efetiva do usuário com o sistema de busca (FERNEDA, 2009).

A aplicação de algoritmos genéticos no campo da recuperação de informação ainda se mostra como uma possibilidade para futuros sistemas com características evolutivas, pois os protótipos existentes são somente testes, que não determinam aplicabilidade em sistemas reais (GORDON, 1988; VRAJITORU, 2000 apud FERNEDA, 2012).

3 DA WEB À WEB SEMÂNTICA

Partindo de seu princípio, a Web (termo utilizado para se referir à WWW – *World Wide Web* – a rede mundial de computadores) foi idealizada a partir dos conceitos de hipertexto e hipermídia, propostos no projeto XANADU, de Ted Nelson, no ano de 1960. Tim Berners-Lee, por sua vez, no ano de 1989, uniu os conceitos de Nelson com a internet, criando então a Web como uma plataforma, inicialmente, com fins acadêmicos. Com seus avanços, e a partir da popularização dos computadores pessoais, tornou-se possível a utilização desta imensa rede como um meio de comunicação, no qual o usuário passa a ter acesso a conteúdos e até mesmo passa a criar páginas, *a priori*, somente informativas.

A Web deve ser entendida como parte da internet e não como um sinônimo desta. Em sua primeira fase, se caracterizava por conteúdos e usos voltados a especialistas, com páginas estáticas. Em seu segundo momento, que teve início por volta de 2004, foi cunhado o termo “web 2.0”, por Tim O’Reilly (2005). É a atual fase da Web, de caráter social e interativo, composta pelos mais diversos ambientes virtuais, representados por *sites* de buscas e compartilhamento de arquivos, blogs, microblogs, redes sociais, wikis, dentre outros serviços que passaram a permitir que o próprio usuário tivesse a possibilidade de disponibilizar informações (pessoais ou não) e interagir com outras pessoas e conteúdos por meio da rede.

A Web Semântica, por sua vez, se caracteriza por ser uma nova fase da Web na qual as informações dispersas na internet são semanticamente descritas de modo a serem recuperadas com maior eficiência e relevância por motores de busca. A proposta de atribuir significado às páginas Web para que sejam interpretadas por

máquinas surgiu no ano de 2001, a partir da publicação de um artigo na revista americana *Scientific American*, cujo título é: "Web Semântica: um novo formato de conteúdo para a Web com significado para computadores vai iniciar uma revolução de novas possibilidades." Tal artigo foi publicado por Tim Berners-Lee (diretor do W3C e pesquisador do Instituto de Tecnologias de Massachusett - MIT), James Hendler (professor da Universidade de Maryland) e Ora Lassila (pesquisador e membro do W3C), e é até hoje um texto referência sobre Web semântica, pois nele os autores definem seus principais conceitos, estrutura e ilustram as situações que esse novo momento da Web pode propiciar (BERNERS-LEE; HENDLER; LASSILA, 2001).

De acordo com Breitman (2005, p. 5), "a ideia central é categorizar a informação de maneira padronizada, facilitando seu acesso." Tais categorias seriam semelhantes a classificações e taxonomias, utilizadas, por exemplo, por biólogos para classificar os seres vivos, classificações que fossem criadas e compartilhadas por diversos pesquisadores do mundo todo, na intenção de estabelecer um modelo estruturado para organizar a desordem informacional da internet.

A atribuição de características semânticas aos conteúdos, de forma que as máquinas possam interpretá-los, leva-se a refletir sobre como esse processo poderá ocorrer. Para que a Web Semântica seja possível, é necessário que diversas ferramentas tecnológicas trabalhem de forma integrada em sua estrutura de implementação. Tais ferramentas podem ser resumidas em: metadados, linguagens de marcação, arquitetura de metadados, ontologias e agentes inteligentes (JORENTE; SANTOS; VIDOTTI, 2009).

As linguagens de marcação são destinadas a estruturar os recursos, de modo a garantir-lhes maior extensibilidade e flexibilidade. As linguagem HTML e XML são exemplos de linguagens de marcação.

O termo ontologia, no contexto da Web, representa um "documento ou arquivo que define formalmente as relações entre termos e conceitos." (SOUZA; ALVARENGA, 2004, p. 137). As ontologias são as estruturas responsáveis pela definição semântica dos conceitos representados pelos metadados, já os agentes inteligentes permitirão a recuperação eficiente dos recursos semanticamente conceituados e devidamente descritos. As arquiteturas de metadados, por sua vez, são representadas por padrões de metadados e linguagens para a representação de

ontologias. Desta forma, são responsáveis pela interoperabilidade dos dados, nos níveis sintático, semântico e estrutural (JORENTE; SANTOS; VIDOTTI, 2009).

Para Breitman (2005, p. 7), o objetivo da Web Semântica é "permitir que máquinas façam o processamento que atualmente [...] tem de ser realizado por seres humanos." A autora enfatiza que os agentes da Web Semântica não substituirão as pessoas, pois estes não tomarão decisões. Sua função será a de "reunir, organizar, selecionar e apresentar as informações a um usuário humano, que tomará suas decisões." (BREITMAN, 2005, p. 8).

Souza e Alvarenga (2004), afirmam que o projeto da Web Semântica se pauta no desenvolvimento e implantação de padrões tecnológicos que facilitem a interoperabilidade de informações entre agentes pessoais e que, sobretudo, estabeleça uma linguagem apropriada para o compartilhamento de dados entre sistemas de informação.

Segundo Castro e Santos (2007), os metadados, no âmbito da Web Semântica, garantem as formas dos recursos informacionais, e as ontologias determinam semanticamente seus conceitos.

A atribuição de descrição semântica aos recursos informacionais possibilita uma recuperação mais eficiente no contexto digital, permitindo, então, que as buscas dirigidas a um sistema por meio de um mecanismo de busca obtenham respostas mais relevantes ao interesse de seu usuário, levando em conta não somente a agilidade da pesquisa e a quantidade de recursos recuperados, mas também a qualidade informacional desses recursos e sua relevância a quem necessita deles. Tais aspectos serão abordados no tópico seguinte.

4 RECUPERAÇÃO NA WEB: DISCUTINDO A QUESTÃO DA RELEVÂNCIA

Como já mencionado anteriormente, os modelos clássicos não são totalmente aptos a serem úteis no ambiente informacional da Web por terem sido desenvolvidos em ambientes fechados, nos quais todo o universo documental é restrito e pode ser conhecido pelo desenvolvedor. Isso não ocorre na Web devido a sua infinita dimensão e capacidade, que permite que, a todo instante, novos documentos sejam adicionados e criados, nos mais variados âmbitos e formatos.

A questão da relevância na recuperação de informação é um tema muito abordado e estudado, porém muito pouco compreendido. Isso se deve ao seu caráter abstrato, que torna difícil criar estruturas artificiais capazes de garantir que os resultados de uma busca sejam relevantes ao seu usuário. Resume-se, basicamente, em mostrar os resultados possivelmente mais relevantes em forma de ranque (*ranking*), do mais relevante ao menos relevante. Entretanto, o conceito de relevância é subjetivo e inexato, não podendo ser definido por fórmulas matemáticas e implementadas em sistemas computacionais.

De acordo com Mizzaro (1998), a relevância possui quatro dimensões. A primeira dimensão é composta por três entidades: documento (a entidade física que o usuário de um sistema de informação obtém depois de buscar pela informação), substituto (uma representação descritiva do documento) e Informação (entidade abstrata que o usuário recebe ao ler o documento). A segunda dimensão diz respeito à representação do problema do usuário, que aparece quando o usuário entra em contato com o sistema de recuperação. O usuário expressa a sua necessidade por meio de uma representação em linguagem natural do que necessita, formatando-a em seguida em sua questão de busca (termo ou palavras-chave pelos quais o usuário fará sua pesquisa no sistema). Deste modo, a segunda dimensão é formada pelas entidades: RIN (Informação real que necessita), PIN (Informação percebida que necessita), solicitação (*request* - representação em linguagem natural da necessidade) e a questão de busca (*query* - formalização da solicitação). Essa segunda dimensão também evidencia alguns problemas que podem ocorrer, como a dificuldade do usuário em estabelecer claramente sua questão de busca (ou mesmo de saber o que necessita buscar), dificuldades em formalizar a questão de forma que o sistema consiga entendê-la, dentre outros. A terceira dimensão é a dimensão do tempo. Algo (um documento, uma informação) pode ser relevante a um usuário em determinado momento e não ser mais em outro, ou vice-versa. Isso pode ocorrer quando o usuário aprende algo que lhe permite compreender um documento ou se a necessidade informacional do mesmo mudar, dentre outras situações. A quarta dimensão é a que diz respeito aos componentes de uma busca. São basicamente três: o tópico (assunto que se refere ao interesse do usuário), a tarefa (execução da busca no sistema) e o contexto (inclui tudo que

não é tratado nas duas outras entidades, e que pode afetar a busca do usuário no sistema).

Desde os primeiros modelos de recuperação de informação, já havia a preocupação com a questão da relevância. Um dos mais comentados modelos é o Vetorial, que trazia a interação do usuário, quando este atribuía os pesos aos termos de sua expressão de busca. A ideia de *relevance feedback* é a de envolver o usuário de um sistema no processo de recuperação das informações, de modo que este possa contribuir para melhores resultados em suas buscas.

Usualmente, em sistemas que utilizam algoritmos de *relevance feedback,* o usuário envia as informações de realimentação ao sistema nos primeiros conjuntos de documentos recuperados. Deste modo, o procedimento básico é: o usuário submete uma expressão de busca ao sistema; o sistema retorna um primeiro conjunto de resultados; o usuário marca alguns dos documentos, os quais julga relevantes à sua expressão de busca; o sistema calcula uma representação das informações baseado nas marcações feitas pelo usuário; o sistema retorna outros resultados revisados. Este processo pode ser repetido quantas vezes o usuário julgar necessário.

Um algoritmo de *relevance feedback* chamado *Rocchio algorithm* foi criado e popularizado pelo sistema SMART (SALTON; BUCKLEY, 1988) por volta do ano de 1970. Este sistema incorporava informações de *relevance feedback* em um modelo Vetorial, que, por sua vez, maximiza a similaridade com documentos relevantes e diminui com documentos não relevantes (MANNING; RAGHAVAN; SCHÜTZE, 2009). Como já mencionado, a relevância, neste modelo, é medida a partir de pesos atribuídos aos termos de busca.

No modelo Probabilístico aproveitam-se os conceitos do modelo Vetorial para otimizar a questão da *relevance feedback* com base na elaboração de uma classificação, que diz que a probabilidade de um termo aparecer em um documento depende de quão relevante ou não ele é. Tendo em mente que um conjunto de documentos relevantes é um pequeno subconjunto do conjunto maior de todos os documentos, a proposta será apta à resolução (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Em qualquer ambiente informacional, a questão da relevância da informação recuperada por um sistema será relativa ao usuário que dela necessita e no

momento que necessita ao momento de que necessita. Dentre os modelos apresentados neste artigo, os que possuem características que podem ser incorporadas à realidade da Web são somente os mais atuais, considerados mais semânticos ou inteligentes, ou seja, as Redes Neurais e os Algoritmos Genéticos. De acordo com Ferneda (2009):

> A utilização dos algoritmos genéticos na recuperação de informação apresenta-se como uma possibilidade para futuras implementações em sistemas com características evolutivas. Sua aplicação rompe com a rigidez dos modelos puramente matemáticos, reconhecendo a inerente indeterminação do processo de representação dos conteúdos dos documentos.

Acerca da utilização dos algoritmos genéticos na Web, o autor aponta que:

> No contexto atual da Web, cuja dinamicidade muitas vezes não permite uma indexação adequada dos documentos a serem disponibilizados, os algoritmos genéticos poderiam representar uma alternativa, ao permitir que as representações dos documentos se configurem adequadamente ao longo de um período, de acordo com a recuperação desses documentos por grupos de usuários de interesses comuns.

O imenso e crescente volume de informações na Web se multiplica a todo instante, e, proporcionalmente aumenta-se a necessidade de dar o tratamento adequado a todos os documentos que nela se encontram e que serão recuperados (ou passíveis de recuperação). Atualmente, a relevância e o ranqueamento das páginas Web são definidas basicamente de duas maneiras: pela repetição das palavras-chave utilizadas na expressão de busca ocorridas no corpo de um texto ou página; ou pela quantidade de vezes que determinada página foi acessada. Além dessas formas, existem também os links patrocinados, que aparecem em primeiras posições nas buscas, tendo quase sempre a ver com o tema pesquisado.

Em relação à Web Semântica e suas preocupações com a relevância da informação, percebe-se que apesar de ter sido pensada como algo mais amplo e desafiador, é hoje mais vista e estudada sob seu aspecto de aperfeiçoar a recuperação da informação, tendo em vista os grandes estoques de informação armazenados desorganizadamente na Web, proporcionando, por meio de suas

tecnologias, uma recuperação mais eficiente, com resultados mais relevantes aos seus usuários.

Por outro lado, sendo utilizada em ambientes Web restritos, como em *sites* de compras, as tecnologias da Web Semântica proporcionam melhorias na interação do usuário com o sistema de busca. Verifica-se que a relevância de seus resultados são provenientes do conhecimento prévio do perfil de seus usuários e, sobretudo, do conhecimento das necessidades que estes possuem ao acessar determinado *site* ou sistema de informação disponível na Web.

5 CONCLUSÕES

Em bases de dados específicas, como bases de dados e catálogos de bibliotecas ou centros de documentação, o processo de recuperação de informação poder ser facilmente desenvolvido pautado em um dos modelos clássicos de recuperação de informação. Isso se deve ao fato de todo o universo ser conhecido e estar acessível seguindo os mesmos padrões. Isso, entretanto não ocorre no ambiente Web, onde as informações estão desorganizadas, com representações fracas e até ausentes, seguindo a diversos padrões de metadados, dentre outros fatores.

A relevância, sendo algo subjetivo e inerente ao julgamento do usuário, dependerá da interação do mesmo com o sistema e, principalmente, ao que de fato ele espera recuperar em sua busca. Apesar da evolução constante dos modelos de recuperação de informação, como apresentado nesse artigo, os sistemas que se baseiam em um modelo de *relevance feedback* não são populares com usuários, uma vez que estes normalmente relutam em fornecer uma resposta explícita ao sistema ou simplesmente não querem prolongar a interação com o mecanismo de busca. Outros problemas podem ser evidenciados pela expressão de busca, como, por exemplo, o uso de expressões longas, compostas por muitos termos.

A utilização das tecnologias da Web Semântica na Web como um todo é, hoje, algo inviável, pois ainda não se mostrou possível o desenvolvimento de ontologias tão avançadas que pudessem dar conta de todos os campos do conhecimento existentes, passíveis de classificar todos os termos e conceitos existentes nos mais variados tipos de recursos informacionais ou que pudessem

comunicar-se umas com as outras, de modo a interoperar seus dados. Até então, a implementação efetiva da Web Semântica só se mostrou possível em domínios restritos ou específicos.

Recentemente, o Google (2013) lançou para usuários norte-americanos o projeto *Knowledge Graph*, no qual suas buscas objetivas seriam complementadas com uma barra lateral, na qual mais informações sobre os termos buscados seriam expostas, de modo a mostrar ao usuário informações básicas sobre os termos ou o conceito pesquisado e outros termos ou conceitos relacionados à busca realizada. A ferramenta tem como principal característica sua capacidade de diferenciar com maior propriedade expressões de busca com termos semelhantes entre si. Entretanto, a efetividade de um instrumento de busca como esse em um ambiente tão heterogêneo como a Web dependerá da constante análise, catalogação e indexação das páginas já existentes e das que passam a existir a todo instante.

A iniciativa demonstra a necessidade e a preocupação em organizar as informações disponíveis na Web e, sobretudo, de fazer com que os resultados das buscas se tornem cada vez mais úteis e relevantes aos usuários utilizadores destes sistemas de recuperação de informação.

REFERÊNCIAS

BERNERS-LEE, Tim; HENDER, James; LASSILA, Ora. **The semantic web:** a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: 10 maio 2012.

BREITMAN, Karin. **Web semântica:** a internet do futuro. Rio de Janeiro: LTC, 2005.

CASTRO, Fabiano Ferreira; SANTOS, Plácida Leopoldina V. A. da Costa. Os metadados como instrumentos tecnológicos na padronização e potencialização dos recursos informacionais no âmbito das bibliotecas digitais na era da web semântica. **Informação & Sociedade: Estudos**, João Pessoa, v. 17, n. 2, p. 13-19, maio/ago. 2007.

FERNEDA, Edberto. Aplicando algoritmos genéticos na recuperação da informação. **DataGramaZero - Revista de Ciência da Informação**, Rio de Janeiro, v. 10, n. 1, fev. 2009. Disponível em: <http://www.dgz.org.br/fev09/Art_04.htm>. Acesso em: 10 maio 2012.

_____. **Introdução aos modelos computacionais de recuperação de informação**. Rio de Janeiro: Ciência Moderna, 2012.

_____. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006. Disponível em: <http://www.scielo.br/pdf/ci/v35n1/v35n1a03.pdf>. Acesso em: 12 maio 2012.

JORENTE, Maria José Vicentini; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Quando as Webs se encontram: social e semântica: promessa de uma visão realizada? **Informação & Informação**, Londrina, v. 14, n. esp., p. 1-24, 2009.

GOOGLE. **Knowledge graph.** Disponível em: <http://www.google.com/insidesearch/features/search/knowledge.html> Acesso em: 20 maio 2012.

KURAMOTO, Hélio. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero - Revista de Ciência da Informação***,* Rio de Janeiro, v. 3, n. 1, fev. 2002. Disponível em: <http://www.dgz.org.br/fev02/Art_03.htm> Acesso em: 16 abr. 2012

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Relevance feedback and query expansion. In: _____. **An introduction to information retrieval.** England: Cambridge University Press, 2009. p. 177-194.

MIZZARO, Stefano. How many relevances in information retrieval? **Interacting with Computers**, Oxford, v. 10, n. 3, jun.1998. p. 303-320.

O'REILLY, Tim. **What Is Web 2.0:** design patterns and business models for the next generation of software. 2005. Disponível em: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> Acesso em: 22 mar. 2012.

PERES, Sarajane Marques; BOSCARIOLI, Clodis. Sistemas gerenciadores de banco de dados relacionais Fuzzy: uma aplicação em recuperação de informação. **Acta Scientiarum**, Maringá, v. 24, n. 6, p. 1733-1743, 2002.

SALTON, Gerald; BUCKLEY, Christopher. Term-weight approaches in automatic text retrieval. **Information Processing & Menagement**, Oxford, v. 24, n. 5, p. 513-523, 1988.

SANT'ANA, Ricardo César Gonçalves. A importância do papel do profissional da ciência da informação nos processos de recuperação de conteúdos digitais estruturados. In: GUIMARÃES, José Augusto Chaves; FUJITA, Mariângela Spotti Lopes (Org.). **Ensino e pesquisa em biblioteconomia no Brasil:** a emergência de um novo olhar. Marília: Cultura acadêmica, 2008. p. 145-154.

SHAW, Ian S.; SIMÕES, Marcelo Godoy. **Controle e modelagem Fuzzy**. São Paulo: Edgard Blücher, 1999.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 2, dez. 2006. p. 161-173. Disponível em: <http://www.scielo.br/pdf/pci/v11n2/v11n2a02.pdf> Acesso em: 12 maio 2012

SOUZA, Renato Rocha; ALVARENGA; Lídia. A Web semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

Title

Information retrieval models and the semantic web: the question of relevance

Abstract

Introduction: In the Web environment, there is a need for greater care with regard to the processing of descriptive and thematic information. The concern with the recovery of information in computer systems precedes the development of the first personal computers. Models of information retrieval have been and are today widely used in databases specific to a field whose scope is known.
Objectives: Verify how the issue of relevance is treated in the main computer models of information retrieval and, especially, as the issue is addressed in the future of the Web, the called Semantic Web.
Methodology: Bibliographical research.
Results: In the classical models studied here, it was realized that the main concern is retrieving documents whose description is closest to the search expression used by the user, which does not necessarily imply that this really needs. In semantic retrieval is the use of ontologies, feature that extends the user's search for a wider range of possible relevant options.
Conclusions: The relevance is a subjective judgment and inherent to the user, it will depend on the interaction with the system and especially the fact that he expects to recover in your search. Systems that are based on a model of relevance are not popular, because it requires greater interaction and depend on the user's disposal. The Semantic Web is so far the initiative more efficient in the case of information retrieval in the digital environment.

Keywords: Information and Technology. Information retrieval models. Semantic Web. Relevance.

Título

Los modelos de recuperación de la información y la web semántica: la cuestión de la pertinência

Resumen

Introducción: La preocupación con la recuperación de la información en los sistemas informáticos precede al desarrollo de los primeros ordenadores personales. Los modelos de recuperación de información han sido y son hoy ampliamente utilizados en bases de datos

específicas de un campo cuyo alcance se conoce. En entorno de la Web, existe una necesidad de un mayor cuidado con respecto al tratamiento de información descriptiva y temática.
Objetivos: Evaluar la forma en que la cuestión de la pertinencia se trata en los modelos de computadoras principales de la recuperación de la información y, sobre todo, ya que el problema se soluciona en el futuro de la Web, llamada de Web Semántica.
Metodología: Investigación bibliográfica.
Resultados: En los modelos clásicos estudiados aquí, se dio cuenta de que la principal preocupación es la recuperación de los documentos cuya descripción es la más cercana a la expresión de búsqueda utilizada por el usuario, lo cual no implica necesariamente que esto realmente necesita. En la recuperación semántica es el uso de ontologías, característica que se extiende la búsqueda del usuario para una amplia gama de posibles opciones relevantes.
Conclusiones: La relevancia es un juicio subjetivo e inherente para el usuario, que dependerá de la interacción con el sistema y, especialmente, el hecho de que se espera recuperar en su búsqueda. Los sistemas que se basan en un modelo de relevancia no son populares, ya que requiere una mayor interacción y dependerá de la voluntad del usuario. La Web Semántica es hasta ahora la iniciativa más eficiente cuando se trata de recuperación de información en el entorno digital.

Palabras clave: Información y Tecnología. Modelos de recuperación de información. Web Semántica. Pertinencia.

Recebido em: 28.06.2013
Aceito em: 10.08.2013