MÁQUINAS DE BUSCA NA WEB E O NOVO PARADIGMA DA WEB SEMÂNTICA

ALEX BRAHA STOLL[1]
ANDREA MASSUMI YAMASHITA[1]
FRANCISCO EDUARDO ALBERTO DE SIQUEIRA GARCIA[1]

RESUMO

O artigo apresenta um estudo dos motivos para a pouca eficácia na recuperação de informações, no paradigma atual da Web. Neste contexto, explica também o funcionamento das search engines, principais ferramentas utilizadas nos dias de hoje para a recuperação de informações na Web. Além disso, ao explorar o novo paradigma da Web Semântica, o artigo apresenta as soluções que têm sido encontradas para melhorar o compartilhamento de informações, por agentes humanos ou não, visando à transformação da Web em um grande sistema de recuperação de informações. Como conclusão, é apresentado o avanço que a Web Semântica proporcionará em diversos campos da área de Tecnologia da Informação.

Palavras-chave: sistemas de recuperação de informação, máquinas de busca, web semântica.

ABSTRACT

The article presents a study of the reasons for the lack of effectiveness in information retrieval, considering the current paradigm of the Web. In this context, it also explains the operation of search engines, the most popular tools for retrieving information on the Web nowadays. In addition, the article presents the solutions that have been found to improve information sharing, by human agents or not, aiming the transformation of the Web into a huge information retrieval system. In conclusion, it is presented the progress that Semantic Web is going to provide in many fields of Information Technology.

Keywords: information retrieval systems, search engines, semantic web.

[1] Graduados, Tecnologia em Análise e Desenvolvimento de Sistemas - Faculdade de Tecnologia de São Paulo, Mogi das Cruzes - SP. e-mail: alexbraha_stoll@gmail.com

INTRODUÇÃO

Vivemos na Era da Informação. Isso significa que esse bem intangível, de características muito particulares, tem muito valor em nossa sociedade. Hoje, deter informação significa ter, também, poder social e econômico.

Dito isto, imediatamente fica clara a importância da Tecnologia da Informação e, mais especificamente, da área de Sistemas de Informação. Esta, como o próprio nome sugere, tem como alvo a sistematização da informação, ou seja, estuda as melhores formas de organizar e disponibilizar a informação para seus usuários.

Tendo isso em mente, fica impossível não falarmos sobre a Web[2]. A rede mundial de computadores, hoje com cerca de vinte e quatro bilhões de páginas[3], é a maior ferramenta para compartilhamento de informações. Tendo acesso à grande rede, qualquer um em qualquer lugar do mundo pode se expressar. Mesmo em países em que há censura, a Web tem se mostrado um mecanismo possibilitador da liberdade de expressão.

Devido a toda essa liberdade de expressão garantida aos usuários da grande rede, todavia, surge um impasse: como disponibilizar e recuperar as informações desejadas dentro de um ambiente que, por não ter controle, acabou se tornando gigantesco e caótico? A área de Sistemas de Informação deu a resposta e, prontamente, a Tecnologia da Informação a implementou na Web.

Recuperar informações é, em outras palavras, utilizar-se de um sistema de recuperação de informações. Estes, na grande rede, são as conhecidas máquinas de busca (search engines). Entretanto, estas máquinas - nada mais nada menos do que algoritmos de busca - nem sempre são capazes de satisfazer as necessidades informacionais dos usuários, pois não são capazes de avaliar corretamente o significado de uma busca.

Surge, então, o paradigma da Web Semântica, com o objetivo de permitir que as máquinas passem a compreender o contexto das consultas dos usuários e, dessa forma, retornem resultados de maior atinência. Portanto, para se estudar o fenômeno da informação em nossa era, é indispensável um detalhado estudo não só da Web mas, também, dos algoritmos que possibilitam a recuperação de informações (as

2 World Wide Web (Grande Rede Mundial), popularmente conhecida como internet, é a gigantesca rede de computadores por meio da qual usuários de todo o mundo interagem, compartilhando informações e experiências.
3 Dado obtido do seguinte endereço eletrônico: <http://www.worldwidewebsize.com>. Acesso em: junho de 2010.

máquinas de busca) e as soluções que têm aparecido para melhorar a disponibilização e recuperação de informações na Web (o paradigma semântico).

1 SISTEMAS DE INFORMAÇÃO

1.1 Conceitos Elementares

Para o estudo dos sistemas de informação, é necessário, antes, compreender conceitos elementares como dado, informação e sistema. Dado é uma descrição elementar, um valor bruto observado e armazenado, incapaz de transmitir qualquer significado.

Coletados os dados e tendo sido organizados e classificados, a ponto de serem valiosos e transmitirem significado a um receptor, passam a ter o status de informação. Estes componentes, em nosso contexto, irão interagir em um sistema, ou seja, um conjunto de partes integrantes e interdependentes, que formam um todo unitário com determinado objetivo e função (ARAÚJO, 1995).

Sistemas de informação são conjuntos de dados interligados que foram organizados e classificados de forma a terem valor e significado para o receptor. Nestes sistemas, é comum primeiro acontecerem atividades que geram os dados e, em seguida, sua coleta e armazenamento; posteriormente, os dados são transformados ou selecionados por meio de um processo, a fim de ganharem significado. Assim, são geradas informações que serão disponibilizadas aos usuários, muito possivelmente ajudando em um processo de tomada de decisão.

1.2 Componentes de um Sistema de Informação

Apesar dos sistemas de informação terem surgido antes da computação, hoje não há como se falar em um sistema desse tipo sem se considerar uma arquitetura de tecnologia que lhe dê suporte. O motivo disto são os benefícios do uso de computadores: maior agilidade na realização de processos, além da maior eficiência na obtenção e transmissão de dados e informações. Posto isto, podemos distinguir os seguintes componentes em um sistema de informação moderno:

- **Hardware:** dispositivos eletrônicos que recebem dados e informações, os processam e os exibem.
- **Software:** conjuntos de programas de computador que servem de interface entre o usuário e o hardware.
- **Banco de Dados:** conjunto de arquivos integrados, para atenderem a sistemas e usuários (HEUSER, 2004).
- **Rede:** sistema de conexão entre dispositivos eletrônicos, permitindo o compartilhamento de recursos.
- **Procedimentos:** as instruções sobre como combinar os componentes para processarem dados e informações, de forma a gerarem os resultados esperados.
- **Usuários:** ao interagirem com hardware e software, usufruem dos recursos oferecidos pelo sistema.

1.3 Sistemas de Recuperação de Informação

Conhecidos os conceitos elementares para a ciência da informação e compreendidos os componentes dos sistemas de informação, temos o cenário necessário ao entendimento de um tipo específico de sistema: os sistemas de recuperação de informação (SRI).

Basicamente, são a interface entre um conjunto de informações e uma comunidade de usuários. Os SRI desempenham as seguintes tarefas: aquisição e armazenamento de documentos; controle e organização dos documentos; e, também, sua disseminação aos usuários (LANCASTER & WARNER apud SOUZA, 2006).

Na Web, os sistemas de recuperação de informações aparecem principalmente como as máquinas de busca (search engines). No universo gigantesco e caótico da grande rede, são os responsáveis pela recuperação das informações solicitadas por usuários de todo o mundo.

2. A WEB COMO SRI: MÁQUINAS DE BUSCA

A Web é, sem dúvida, o maior repositório de informações existente. Com cerca de vinte e quatro bilhões de páginas[4], contém, senão todo, grande parte do conhecimento já produzido pela humanidade. Praticamente toda essa informação,

[4] Dado obtido do seguinte endereço eletrônico: <http://www.worldwidewebsize.com>. Acesso em: junho de 2010.

entretanto, está dispersa sem qualquer critério de organização. Disso, surge um impasse: um caótico e colossal repositório de informações de um lado; do outro, os usuários desse sistema, que precisam recuperar informações específicas em meio à imensidão da Web. Na tentativa de solucionar essa problemática, há os **sistemas de recuperação de informações** (SRI), popularmente conhecidos na Web como search engines (máquinas de busca).

Atender às necessidades do usuário, porém, não é tarefa fácil. Os sistemas de informação carregam complicações inerentes ao conceito de informação (ARAÚJO, 1995). Diferentemente dos **sistemas de gestão de bancos de dados** (SGDB), não é possível nos SRI a obtenção de uma resposta exata. No primeiro tipo, há a manipulação de dados organizados matricialmente; no segundo sistema, todavia, o conteúdo é a informação, que não pode ser separada do usuário, pois surge no momento da interpretação. Assim, em SRIs é difícil se determinar a real necessidade do usuário, pois há fraca associação entre os registros do acervo e seus conteúdos informativos (SOUZA, 2006).

Para se estudar a qualidade da recuperação de informações e melhor compreender as principais metodologias utilizadas pelas search engines, é necessário o conhecimento de duas medidas dos resultados de uma busca: **revocação** e **precisão**. A primeira, conhecida também como recall, é a razão entre a quantidade de documentos atinentes recuperados e a totalidade de documentos atinentes disponíveis, indicando a competência do SRI na recuperação de documentos pertinentes. Precisão, por sua vez, é a razão entre a quantidade de documentos pertinentes recuperados e a totalidade de documentos resgatados, indicando o quão bom o SRI é em não recuperar documentos incoerentes às necessidades de informação do usuário.

O problema central em SRIs é a separação entre os documentos que são relevantes, de acordo com as necessidades do usuário, e aqueles que devem ser descartados. Quando tratamos de search engines, essa escolha é feita por algoritmos, que também ordenam os documentos por relevância baseando-se em critérios previamente estabelecidos (BAEZA-YATES e RIBEIRO-NETO apud SOUZA, 2006). Podem-se dividir os modelos de recuperação em dois tipos: clássicos e estruturados. Naqueles, o assunto e conteúdo de cada documento são representados por palavras-chave. Nos modelos estruturados, podem-se especificar não só palavras-chave, mas também informações quanto à estrutura do texto, como seções a serem pesquisadas,

fontes de letras, proximidade entre palavras etc (SOUZA, 2006). Dos algoritmos tipo clássico, podemos destacar três: o modelo **booleano**, o modelo **vetorial** e o modelo **probabilístico**. A seguir, esses modelos como são apresentados por Baeza-Yates e Ribeiro-Neto apud Souza (2006):

> **Modelo booleano:** é uma solução simples e elegante, baseada na teoria dos conjuntos. Longe de ser a mais eficaz, esse algoritmo recupera todos os documentos que possuam as palavras-chave determinadas pelo usuário, que pode ainda relacioná-las utilizando os operadores booleanos (or, and e not). Esse modelo possui como pontos fracos a classificação dualista dos documentos (relevante ou irrelevante) e a ausência de capacidade de ordenação dos resultados da busca.
>
> **Modelo vetorial:** é o algoritmo base da grande maioria dos SRIs, sendo amplamente utilizado em search engines. Neste modelo, os documentos são representados como vetores em um espaço n-dimensional, em que n é a totalidade de palavras-chave de todos os documentos no sistema. Como não é um algoritmo dualista, é possível serem determinados graus de relevância para os documentos recuperados, construindo-se um ranking. As search engines utilizam também outras técnicas para a ordenação, que serão estudadas posteriormente neste artigo.
>
> **Modelo probabilístico:** neste tipo de algoritmo, supõe-se que, para cada consulta ao sistema, há um conjunto ideal de documentos que a satisfaça completamente. Por meio de tentativa inicial com uma coleção de documentos (para a qual se podem usar técnicas de outros modelos, como o vetorial) e do feedback do usuário em sucessivas interações, busca-se uma aproximação do conjunto ideal. O valor desse modelo é dar grande importância às respostas do usuário como forma de aprimoramento contínuo das buscas.

Os modelos descritos são apenas uma amostra dos algoritmos que vem sendo utilizados para a recuperação de informações. Hoje, é consenso a necessidade de pesquisas em diversas frentes, para melhor atender às necessidades informacionais dos usuários.

Por mais sofisticados que sejam os algoritmos existentes, fica cada vez mais claro que eles não são suficientes (se considerarmos como objetivo a melhoria contínua) em uma Web com cerca de vinte e quatro bilhões de páginas[4]. O motivo disso é que tais algoritmos utilizam palavras isoladas para a busca, não considerando o contexto informacional implícito em toda consulta. Como é esse contexto o determinante das minúcias e especificidades do assunto pesquisado, perdem-se informações fundamentais sobre o sentido em que os termos estão sendo utilizados, diminuindo em muito a pertinência dos resultados da consulta (SOUZA, 2006).

Solucionar esse problema exige, então, muito mais do que a criação de novos algoritmos: na verdade, é necessária uma mudança no paradigma da Web. Essa revolução - que já vem acontecendo - chama-se **Web Semântica**. Além da exploração do sentido intrínseco de qualquer consulta, propõe também a criação de padrões de metadados (dados sobre as próprias páginas Web), de forma que cada site possua uma descrição fortemente contextualizada de seu conteúdo, permitindo inclusive o entendimento desse contexto por programas de computador (comumente chamados de agentes inteligentes). A seguir, explica-se com maiores detalhes o que é o paradigma da Web Semântica e como ele pretende proporcionar aos usuários maior satisfação na recuperação das informações que desejem.

3. O NOVO PARADIGMA DA WEB SEMÂNTICA

Segundo Berners-Lee (2001), a Web Semântica deve ser vista como uma extensão da Web atual. A principal característica dessa extensão é dar um claro significado às informações, melhorando a interação entre pessoas e máquinas.

A Web Semântica consiste, então, na criação e implantação de padrões tecnológicos para facilitar a troca de informações entre pessoas e, também, na confecção de uma linguagem para o compartilhamento mais eficaz de dados entre agentes não humanos. Com todos os usuários da Web seguindo padrões na descrição e armazenamento dos dados, será possível, pois, a utilização da informação de maneira automática e não ambígua, seja por agentes humanos ou não (SOUZA e ALVARENGA, 2004).

Como exemplo de uma das principais tecnologias para a concretização da Web Semântica, se pode citar o XML (Extensible Markup Language). Enquanto a

[4] Dado obtido do seguinte endereço eletrônico: <http://www.worldwidewebsize.com>. Acesso em: junho de 2010.

linguagem HTML (Hypertext Markup Language) concentra-se no controle da forma como serão exibidos os dados, o XML tem como foco descrever os dados que possui um documento. Somente o XML, todavia, não é suficiente para construir metadados: é necessário, também, um padrão para a comunicação, de forma que esta seja consensual, inteligível e não ambígua para todos os participantes de uma comunidade (SOUZA e ALVARENGA, 2004).

3.1 O Padrão RDF

O RDF (Resource Description Framework) é uma das tecnologias que estão sendo utilizadas na confecção de páginas da Web Semântica. Trata-se da definição de um padrão de metadados para ser embutido em codificação XML. O RDF permitirá que, utilizando-se a infra-estrutura do XML, haja na Web um ambiente consistente para publicação e utilização de metadados. Além disso, proverá uma sintaxe padronizada para a descrição do conteúdo e propriedades dos documentos na Web. Como consequência de todos esses benefícios, permitirá que se aja de forma inteligente e automatizada sobre as informações na Web, uma vez que seus significados se tornarão mais facilmente compreensíveis (SOUZA e ALVARENGA, 2004).

Apesar da grande valia do RDF, ainda há o que melhorar: é preciso encontrar maneiras para que a descrição dos namespaces (definição de vocabulários controlados que identificam um conjunto de conceitos de forma única) seja mais inteligente, não repetitiva e compreenda mais propriedades (SOUZA e ALVARENGA, 2004). Neste âmbito, a seguir será estudado ontologia, um tipo mais genérico de namespace.

3.2 Ontologia

Segundo Souza e Alvarenga (2004), "a palavra ontologia deriva do grego *onto* (ser) e *logia* (discurso escrito ou falado)". As ontologias, para a ciência da informação, são modelos compartilhados para a definição de conceitos, propriedades e axiomas, modelos estes legíveis pelo computador (SILVA; SOUZA; ALMEIDA, 2008). Resumidamente, então, ontologias são

documentos ou arquivos nos quais as relações entre termos e conceitos são definidas formalmente (PICKLER, 2007).

Construir uma ontologia, pois, é buscar satisfazer a necessidade de troca de informações entre os membros de uma comunidade, sejam eles humanos ou não. Para se construir esse vocabulário compartilhado, estão sendo criados padrões e linguagens, todos baseados no XML (SOUZA e ALVARENGA, 2004).

3.3 Agentes Inteligentes

Podem-se definir agentes inteligentes como programas que, por meio do uso da inteligência artificial, auxiliam o usuário na realização de tarefas. Como já dito anteriormente, o sucesso da Web Semântica depende da manipulação e processamento automático e autônomo de informações. Os agentes inteligentes serão os responsáveis por isso, coletando a informação de diversas fontes e a processando para, enfim, disponibilizá-la ao usuário humano (SOUZA e ALVARENGA, 2004).

Segundo Wooldridge & Jennings apud Souza e Alvarenga (2004), os agentes inteligentes terão autonomia, serão capazes de interagir com outros agentes (artificiais ou humanos) e serão capazes de reagir a mudanças no ambiente. Além disso, esses programas - que estarão continuamente sendo executados - captarão informações sobre o usuário e o ambiente, fazendo com que possam agir de maneira proativa e sejam orientados a objetivos.

Por fim, é importante ressaltar que o avanço dos agentes inteligentes acompanhará o desenvolvimento da Web Semântica, pois maior quantidade de informações será "compreensível" para a máquina. Assim, haverá uma mudança na relação entre o homem e o computador, com o aumento da quantidade de tarefas delegadas a máquina (SOUZA e ALVARENGA, 2004).

3.4 O Avanço que a Web Semântica Proporcionará

O novo padrão da Web Semântica proporcionará uma série de

melhorias no que diz respeito à recuperação de informações. Algumas das mais importantes, de acordo com Souza e Alvarenga (2004), são:

> **Melhores motores de busca -** as metodologias de marcação semântica dos dados (representadas pelos metadados e namespaces da Web Semântica) e a lógica formalizada do XML e do RDF prometem melhorar drasticamente a recuperação de informações pelas search engines.
>
> **Construção de novas interfaces com o usuário** - a lógica intuitiva e natural do RDF permite que projetemos interfaces para sistemas de informação de uma forma mais coerente com o funcionamento cognitivo dos seres humanos. Além disso, com os agentes inteligentes, é possível aprimorar e personalizar a utilização dos perfis de usuários para que a interação destes com os sistemas seja mais significativa e ágil.
>
> **Gestão do conhecimento organizacional -** Com as ontologias comunitárias e a padronização dos metadados, explicitar, classificar e armazenar o conhecimento produzido dentro das organizações se tornará muito mais fácil. Assim, deve-se esperar grande avanço dos portais corporativos, que são a principal tecnologia quando o assunto é gestão do conhecimento.

CONCLUSÃO

Atualmente, falar em compartilhamento de informações é falar em Web. Esta grande rede é um repositório gigantesco e caótico de informações. Por isso, para termos nossas necessidades informacionais satisfeitas, é necessário fazer uso dos sistemas de recuperação de informação, que na Web são as conhecidas máquinas de busca.

O modelo atual das search engines, todavia, não atende satisfatoriamente às requisições dos usuários, pois é incapaz de compreender o contexto implícito presente nas consultas realizadas. Para solucionar esse impasse, a Web está sendo aperfeiçoada: trata-se do paradigma da Web Semântica que, basicamente, tem como objetivo fazer com que os algoritmos de busca sejam capazes, também, de compreender o significado das

requisições dos usuários.

Concluí-se, portanto, que esse novo paradigma irá proporcionar drástica melhoria na recuperação de informações, permitindo o aperfeiçoamento dos motores de busca e das interfaces dos SRI, além de melhor gestão do conhecimento no ambiente organizacional. Assim, é esperado para os próximos anos um desenvolvimento na troca de informações, transformando a Web em um ambiente ainda mais propício ao florescimento cultural e científico.

REFERÊNCIAS

ARAÚJO, Vânia M.R.H. **Sistemas de recuperação da informação: nova abordagem teórico conceitual**. 1994. Tese (Doutorado em Ciência da Informação). Universidade Federal do Rio de Janeiro, Rio de Janeiro. Disponível em: <http://dici.ibict.br/archive/00000141/01/Ci%5B1%5D.Inf-2004-577.pdf>. Acesso em: abril de 2010.

BERNERS-LEE, T. *et al*. **The semantic toolbox: building semantics on top of XML -RDF**. Disponível em: <http://www.w3.org/ DesignIssues/Toolbox.html>. Acesso em: abril de 2010.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados.** Porto Alegre: Sagra-Luzzato, 2004.

PICKLER, Maria Elisa Valentim. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspect**. Ciênc. Inf., Belo Horizonte, v. 12, n. 1, apr. 2007 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100006&lng=en&nrm=iso>. Acesso em: maio 2010.

SEMANTIC web. Disponível em: <http://www.semanticweb.org/about.html>. Acesso em: maio de 2010.

SILVA, Daniela Lucas da; SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos. **Ontologias e vocabulários controlados: comparação de metodologias para construção.** Ci. Inf., Brasília, v. 37, n. 3, Dec. 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652008000300005&lng=en&nrm=iso>. Acesso em: maio 2010.

SOUZA, Renato Rocha; ALVARENGA, Lídia. **A Web Semântica e suas contribuições para a ciência da informação.** Ci. Inf., Brasília, v. 33, n. 1, apr. 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100

19652004000100016&lng=en&nrm=iso>. Acesso em: maio 2010.

SOUZA, Renato Rocha. **Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. Perspect**. Ciênc. Inf., Belo Horizonte, v. 11, n. 2, aug. 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362006000200002&lng=en&nrm=iso>. Acesso em: maio de 2010.

WORLD wide web size.com. **The size of the World Wide Web (The Internet).** Disponível em: <http://www.worldwidewebsize.com>. Acesso em: jun. 2010.