UMA ANÁLISE DO FUNCIONAMENTO DOS MECANISMOS DE BUSCA NA REDE MUNDIAL DE COMPUTADORES

Por

Donizeti Batista

Orientador: Prof. Dr. Cláudio Bornstein

RIO DE JANEIRO – RJ

MARÇO 2007

UMA ANÁLISE DO FUNCIONAMENTO DOS MECANISMOS DE BUSCA NA REDE MUNDIAL DE COMPUTADORES

Donizeti Batista

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM HISTÓRIA DAS CIÊNCIAS E DAS TÉCNICAS E EPISTEMOLOGIA.

Aprovada por:

_____
Prof. Claudio Thomas Bornstein, Ph.D.

_____
Prof. Ricardo Silva Kubrusly, Ph.D.

_____
Dra. Adriana Lúcia Cerri Triques, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 2007

BATISTA, DONIZETI

Uma análise do funcionamento dos mecanismos de busca na rede mundial de computadores [Rio de janeiro] 2007

IX, 91 p. 29,7 cm (COOPE/UFRJ, M.Sc., História das Ciências e das Técnicas e Epistemologia, 2007)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Ciência, Tecnologia e Sociedade
2. Mecanismos de Buscas na *Web*

I. COPPE/UFRJ II. Título (série)

AGRADECIMENTOS

A hora dos agradecimentos é sempre difícil, pois além dos fantasmas da injustiça e do esquecimento, tem-se o sentimento que todo reconhecimento é pouco aos muitos que de alguma forma acompanharam e apoiaram a execução deste trabalho.

Gostaria de agradecer inicialmente ao HCTE, seus professores, companheiros alunos, e à secretária Lúcia.

Ao meu orientador Profº Claudio B. por seus incentivos e considerações em suas várias revisões e encaminhamentos à execução deste trabalho

Ao professor Luiz Alfredo pelo seu empenho fundamental para o reconhecimento deste Programa, garantindo uma excelente oportunidade aos seus alunos.

Ao professor Kubrusly, certamente aquele que encarna o espírito deste grupo, ainda mais pelos inusitados encerramentos de seus cursos no Pico da Tijuca e visitas à Praça da Prainha.

À Katinha, que me arrastando pelos corredores do Fundão, mostrou–me que a persistência e o cuidado com os amigos nos levam... a comer muitos salgadinhos!

Aos meus companheiros de turma do HCTE, que muito me ensinaram através das conversas divertidas e discussões acaloradas nas horas do café, nos encontros com músicas e acadêmicos do EAHCTE, estes últimos orquestrados pela queridíssima Virginia, que também com atenção leu os primeiros escritos e deu-me a certeza de que precisei para chegar ao final deste trabalho.

Um agradecimento especial à Dra. Adriana Triques por seu incentivo e anotações antecipadas aos rascunhos deste trabalho.

Não poderia deixar de lembrar DRI, não apenas por seus incentivos e elogios, mas por gastar parte de seu precioso tempo, que andava bastante curto, para ler, revisar, opinar, trinchar, entre outros, no tortuoso período de finalização deste trabalho.

À compreensão e atenção dos meus colegas e amigos da equipe de trabalho no Datasus, principalmente, ao Jonas que me permitia as tantas flexibilidades de horário.

E finalmente aos meus irmãos Claudio e Vicente, companheiros de Rio e de CEU, dando incentivos, puxões de orelhas e metendo a mão na massa...

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA ANÁLISE DO FUNCIONAMENTO DOS MECANISMOS DE BUSCA NA REDE MUNDIAL DE COMPUTADORES

Donizeti Batista

Março 2007

Orientador: Cláudio Bornstein

Programa: História das Ciências e das Técnicas e Epistemologia

A *Web* é uma fonte de informações que vem adquirindo uma crescente influência na formação educacional e sócio-cultural em nível mundial. Neste espaço, as ferramentas de busca de informação passam a exercer um papel fundamental. Por este motivo, é de grande importância o esclarecimento dos critérios técnicos e dos interesses envolvidos na estruturação das diversas ferramentas de busca disponíveis na *Web*.

No presente trabalho, procura-se fomentar uma discussão acerca dos interesses econômicos envolvidos nas práticas empregadas por empresas que disponibilizam mecanismos de busca de informação na internet.

Inicialmente, é feita uma breve descrição do funcionamento de diversos tipos de ferramentas de busca, ressaltando como os métodos empregados pelas empresas que os oferecem podem, arbitrariamente ou não, influenciar o resultado da pesquisa. Também são descritas algumas das relações comerciais que podem ser estabelecidas com empresas que oferecem inclusão e/ou posicionamento de páginas na listagem de resultados obtida com uma determinada ferramenta de busca.

Faz-se um estudo de caso analisando-se de maneira crítica as práticas empregadas pela empresa Google Inc., uma das maiores empresas de mecanismo de busca na internet atualmente. Mostra-se que algumas destas práticas não estão sendo adotadas de maneira clara, o que pode induzir a erros de interpretação dos resultados da pesquisa e a limitações na liberdade dos usuários. A partir desta análise, argumenta-se que os métodos empregados por estas empresas devem ser mais amplamente esclarecidos e discutidos pela sociedade.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.).

UMA ANÁLISE DO FUNCIONAMENTO DOS MECANISMOS DE BUSCA NA REDE MUNDIAL DE COMPUTADORES

Donizeti Batista

March 2007

Advisor: Cláudio Bornstein
Programa: História das Ciências e das Técnicas e Epistemologia

The World Wide Web is an information source with a growing influence on educational and socio-cultural aspects. In this cyberspace, the search engines play a fundamental role. Therefore, it is important to highlight the technical and economical criteria employed by these engines.

First, a general description search engines is made. It is pointed out how the employed mechanism may, arbitrarily or not, influence the result of the research. Also, some of the commercial relationships between Web page owners and search engines are described, examining the mechanisms for the inclusion and ranking of the Web pages.

The methods and procedures used by Google Inc., one of the biggest search engine companies in the world, are examined. It is shown that some of these practices lack of transparency and can even induce to errors in the search results. Based on this analysis, it is argued that an explicit description of the mechanisms employed might be given by the search engine companies and that the methods adopted must be widely discussed by society.

INDICE

ÍNDICE DE FIGURAS

INTRODUÇÃO

As ferramentas de busca de informações na *Web* tornaram-se um elemento de extrema importância na Internet, não apenas pelas suas funcionalidades, mas também pelo que representam para o comércio virtual, um mercado que movimentará até o final da década mais de vinte e cinco bilhões de dólares/ano, segundo projeções[1]. Vale ressaltar que a mais popular destas ferramentas – o Google – teve um faturamento em torno de seis bilhões de dólares em 2005 e estima-se que atinja os nove e meio bilhões de dólares em 2006[2]. É também possível verificar uma forte concentração dos serviços de busca na *Web* entre poucas empresas: segundo pesquisa da Nielsen/Netratings, as três mais populares ferramentas de busca na *Web* – Google, Yahoo! e MSN Search - concentram mais de 90% das consultas realizadas pelos usuários deste serviço.

Por questões históricas, técnicas, e econômicas, que serão detalhadas ao longo deste trabalho, o resultado apresentado ao usuário por uma ferramenta de busca se dá através de um banco de dados que está sob o controle de organizações que oferecem este serviço na *Web*. No momento em que o usuário aciona a ferramenta digitando uma palavra ou expressão, a busca é realizada neste banco de dados, onde as informações foram previamente adicionadas e ordenadas. O usuário não obtém suas informações instantânea e diretamente na *Web*, como induz o raciocínio proposto pelo sistema, mas

---

[1] Segundo John Batelle, co-fundador da revista Wire, jornalista e especializado em ferramentas de busca na *Web*, à revista EXAME (09/11/2005).

[2] A Google foi uma das empresas que mais lucraram com a Internet, no primeiro trimestre de 2006, segundo a comScore (http://www.comscore.com), resultados de julho de 2006.

apenas parte das informações existentes na *Web,* ordenadas e disponibilizadas segundo os recursos técnicos, a lógica econômica, e os interesses dos organizadores ou empresas que oferecem o serviço de ferramenta de busca na *Web*. Cabe destacar que pouca informação é disponibilizada por estas empresas sobre os critérios organizacionais que utilizam e como os manipulam. Essa obscuridade no funcionamento de tais ferramentas é uma questão que merece ser discutida, reforçando a necessidade da abordagem deste tema.

A importância do tema é realçada por Introna e Nissenbaum (2001, p. 181) quando afirmam que é necessário o esclarecimento das regras e critérios de funcionamento e dos algoritmos utilizados nos programas, o que contribui para um melhor entendimento do funcionamento destas ferramentas, apesar dos "riscos" que possam advir daí, como o uso que poderia ser feito por outras empresas ou indivíduos (*spammers*), como alegam os proprietários das ferramentas de busca.

Esclarecer as diferenças existentes entre o discurso das empresas da sua prática ao implementarem seus critérios organizacionais, em nosso entendimento, é de crucial importância para estabelecer as influências dos parceiros comerciais na escolha e ordenamento das páginas apresentadas por estas ferramentas.

Este trabalho se propõe a apresentar as principais características do funcionamento das ferramentas de buscas na Internet e suas implicações na ordenação das informações que são apresentadas aos usuários da *Web*. Discutiremos o que são, como são conhecidas, e qual o funcionamento destas empresas ou grupos que oferecem estes recursos aos usuários, reforçando a importância da discussão do papel das empresas detentoras das ferramentas e sua influência no desenvolvimento e ordenação das informações oferecidas na Internet.

Buscaremos, ainda, tratar das possíveis implicações econômicas e culturais resultantes da forma de estruturação das ferramentas de busca na Internet, questão esta pouco discutida entre pesquisadores e profissionais da área, bem como pela comunidade usuária destes recursos (DIAS, 2005).

Serão usados artigos de especialistas da área e da comunidade acadêmica, que serão citados ao longo do texto, assim como informações divulgadas pelas próprias empresas responsáveis por estas ferramentas.

No primeiro capítulo será feita uma descrição das principais características das ferramentas de busca na *Web* conhecidos e as distinções entre os vários métodos de busca utilizados.

No segundo capítulo se tratará das características particulares do serviço da Google Inc., discutindo algumas limitações do seu mecanismo de busca e contradições nos discursos e posições oficiais tomados por esta empresa.

No capítulo 3 são discutidas questões relativas aos interesses comerciais vinculados aos mecanismos de busca e os interesses econômicos deste setor, destacando como um pequeno grupo de empresas desenvolve e detém o controle destas ferramentas que movimentam bilhões de dólares anuais e geram grandes lucros utilizando-se de um discurso de imparcialidade em relação à coleta e apresentação das informações da Internet. Será destacado como os interesses de parceiros e/ou clientes comerciais das empresas que oferecem as ferramentas de busca podem estar influenciando a apresentação das informações oferecidas ao usuário.

O capítulo 4 traz as considerações finais deste estudo, onde se discutem as vantagens para a sociedade da ampliação da análise e discussão sobre as ferramentas de

busca de informação na *Web*, estrutura que, a cada dia, expande o seu papel neste ciberespaço sócio-político e econômico, que é hoje a Internet.

Capítulo 1

CARACTERÍSTICAS E ASPECTOS FUNCIONAIS DOS MECANISMOS DE BUSCA

É de fundamental importância para este trabalho, explicitar algumas características técnicas e teóricas das ferramentas de busca na *Web* e as estratégias utilizadas na sua implantação. Para que possamos discutir como as decisões ligadas a estas técnicas influenciam os resultados das buscas efetuadas pelos usuários, serão apresentadas neste capítulo as principais características do funcionamento das ferramentas de busca na *Web*.

As ferramentas de busca às quais estaremos fazendo referência são as que atuam *on-line* na *World Wide Web*[3] (*Web*) (BERNERS-LEE, 1997), sejam estas comerciais ou com outros fins.

As descrições dos mecanismos de busca aqui realizadas são de caráter genérico e se reportam aos procedimentos mais comuns utilizados, seja no passado ou atualmente, pelas ferramentas de busca na *Web*. Esta fora do escopo deste trabalho descrever o funcionamento da ferramenta de uma empresa específica. Quando fizermos referência a uma empresa em particular será devido à relevância das informações disponibilizadas pela própria, ou em função do destaque dado a ela pela literatura disponível sobre o assunto[4].

---

[3] Tim Berners-Lee, Director of the World Wide Web Consortium e considerado o criador da *Web.*

[4] Como por exemplo, um artigo acadêmico que foi escrito por um dos co-fundadores da Google Inc. Considerado como a primeira descrição do algoritmo de *PageRank*, patenteado e que é utilizado

São tomadas como base informações que podem ser encontradas em publicações de técnicos, analistas, e pesquisadores da área, bem como dos próprios proprietários ou responsáveis pelas empresas ou organizações desenvolvedoras das ferramentas de busca, assim como manuais que servem de tutorial para a utilização otimizada dos mecanismos de busca.

1.1. Internet e World Wide Web

Antes de iniciarmos a apresentação das ferramentas de busca, avaliamos ser necessário explicitar as diferenças básicas entre os conceitos de Internet e da *World Wide Web.*

É bastante comum na literatura e entre profissionais da área o uso das expressões Internet e *Web* (*World Wide Web,* WWW) como termos intercambiáveis (ANTUNE e CORREIA, 2003). No entanto, não são sinônimos apesar de Internet e *Web* serem estruturas inter-relacionadas.

A Internet é uma rede que conecta redes de computadores (Lévy, 2000). Redes de computadores são sistemas de comunicação entre computadores, que permitem a troca de informações entre eles. No caso específico da Internet, é uma rede que integra ou interliga redes de computadores. Esta comunicação pode ser categorizada por função, sendo que um computador terá a função de servidor e outro terá a função de cliente,

---

como base do funcionamento do mecanismo de busca desta empresa, que será melhor descrito adiante.

onde o primeiro disponibiliza um serviço ou aplicação e o segundo se utilizará destas. Podemos adiantar aqui que, no caso da WWW, é no servidor *Web* que as páginas *Web* são disponibilizadas e é no computador cliente onde as páginas são visualizadas. É certo que, no servidor *Web*, também podemos visualizar páginas, assim este está executando ambas as funções: de servidor e de cliente.

A Internet integra uma grande estrutura na qual estão definidos alguns padrões, principalmente os de comunicação entre computadores, que são conhecidos por Protocolos de Comunicação. As informações são trocadas entres os computadores através de vários destes protocolos, mas a principal característica é a utilização do TCP/IP[5] como base desta comunicação[6]. Alguns serviços disponíveis na Internet utilizam-se de protocolos próprios, mas a transmissão é realizada pelo TCP/IP (*TCP over IP*). Podemos exemplificar algumas dessas aplicações: o correio eletrônico, que utiliza os protocolos SMTP, POP e o IMAP[7]; os serviço de noticias (*NEWS*) com o protocolo NNTP[8], ou mesmo os aplicativos de troca de mensagens (*Messengers*) como ICQ, MSN Messenger, bem como as redes de troca de arquivos KaZaA e Napster, que

---

5 O TCP/IP envolve de fato dois protocolos – o IP (*Internet Protocol*) que é responsável pelo encaminhamento dos dados e o TCP (*Transfer Controller Protocol*) que controla a transferência e possíveis erros na transmissão.

6 Na Camada de Rede (*Network Layer*), fizemos essa nota apenas para evitar confusão, caso um especialista venha a ler este trecho, apesar do objetivo neste momento é apresentar características técnicas para o entendimento de questões que virão a frente.

7 Simple Transfer Mail Protocol (SMTP) que é utilizado no envio e/ou para a transferência de e-mails entre os servidores, Post Office Protocol versão 3 (POP3) e Interactive Mail Access Protocol versão 4 (IMAP4) que são os principais protocolos padrões responsáveis pelo recebimento e acesso remoto a e-mails, respectivamente.

8 Network News Transfer Protocol

são baseados no P2P[9]. Há muitos serviços que têm seus próprios protocolos de troca de informação, isso possivelmente decorre da característica de um espaço virtual aberto que é a Internet propiciando o aparecimento de novas aplicações. No entanto, o nosso objetivo, neste momento, é apenas apresentar algumas características dos serviços disponíveis da Internet.

Se, por um lado, a Internet envolve todas essas aplicações, por outro lado, a *Web* é apenas uma dessas aplicações, que também está disponível nesta estrutura. O HTTP (*HyperText Transfer Protocol*) é o protocolo utilizado na transferência de dados na *Web* - como especificado na RFC 1945[10]:

> "(...) *The Hypertext Transfer Protocol (HTTP) is an application-level protocol with the lightness and speed necessary for distributed, collaborative, hypermedia information systems*" (BERNERS-LEE, FIELDING E FRYSTYK, 1996).

Disponibiliza-se uma camada ou interface multimídia (texto, imagem e som) para o acesso a informações na Internet.

O HTTP, juntamente, com o HTML (*Hypertext Makup Laguage*) - esta uma linguagem de programação de marcação e de referência a documentos na Internet - permitiram a criação da *Web*. As páginas *Web* (*Web Pages)* são documentos que estão codificados na linguagem HTML, em marcações chamadas "*tags"*, mediante as quais

---

9 *Peer-to-Peer Protocol,* permite a troca de informações diretamente entre os computadores dos usuários do serviço, reduzindo o trafego a um servidor que centralizasse as comunicações.

10 As RFC (*Request for Comments*) são publicações de especificações consideradas como padrões para a Internet pela *Internet Engineering Task Force* (IETF), oficialmente responsável por promover e organizar estas normas. Todas as RFC têm uma numeração. No caso a que especifica o HTTP é a de número 1945.

são determinadas as formatações do texto, imagens e demais elementos que compõem estas páginas. A página *Web* pode também ser definida como um documento composto de referências (*links*) para outros documentos. Apesar do conceito de página *Web* ser mais amplo, este caráter de inter-relacionamento entre estas páginas é suficiente para o propósito deste trabalho.

São os programas chamados Navegadores Internet (*Web Browsers*) que possibilitam visualizar as páginas *Web* com seus recursos multimídia e, através de seus *links*, acessar outras páginas e locais (*sites*).

O correio eletrônico não é uma aplicação *Web*, mas quando disponibilizado em página *Web* (*Webmail*), com acesso através de um navegador, podem ser considerado como tal. Desta forma, muitos serviços originalmente disponíveis apenas na Internet, poder ter suas características alteradas, com seus protocolos de comunicação encapsulados no protocolo HTTP, tornando-se uma aplicação integrante da *Web*:

> "HTTP is also used as a generic protocol for communication (...) to other Internet protocols, such as SMTP [12], NNTP [11], FTP [14], Gopher [1], and WAIS [8], allowing basic hypermedia access to resources available from diverse applications and simplifying the implementation (…)". (BERNERS-LEE, FIELDING e FRYSTYK. 1996)

Assim, um usuário de aplicações disponíveis na Internet, pode ter dificuldades em distinguir se está utilizando recursos da *Web* ou da própria Internet. Ou seja, embora *Web* e Internet apresentem linhas divisórias nítidas para um especialista na área, para um usuário comum elas podem não passar de expressões sinônimas.

1.2. Métodos de Acesso a informações na *Web*

Como um usuário iniciante encontraria uma informação na *Web*[11] ? Como descrevem Maze, Moxley e Smith (1997), há três maneiras de fazê-lo. Primeiramente, o usuário poderia, partindo de um *site* indicado por alguém, ou até mesmo do portal do seu próprio provedor de acesso à Internet[12], ir seguindo os *links* em várias direções, ou seja, "navegando" entre esses *links*, de página em página, na *Web*. Utilizando esta estratégia e dispondo de algum tempo, o usuário poderá encontrar um volume grande de informações. No entanto, se ele desejar uma informação muito específica e lhe faltar habilidade, experiência, tempo ou algum dado que permita percorrer os *links* destas páginas, ele poderá não alcançar seu objetivo, tendo sua navegação interrompida.. Para este usuário, este mecanismo ou procedimento pode não ser o mais adequado.

Uma segunda maneira poderia ser buscar as informações utilizando-se um guia de assuntos para a *Web*. Este método pode ser mais eficiente para que nosso usuário hipotético encontre as páginas desejadas. Os "Catálogos de assuntos" ou Diretórios, como são conhecidos, formam a base do funcionamento de diversas ferramentas de busca de empresas ou corporações como Yahoo! Inc. e DMOZ[13] entre outras que podem ser encontradas atualmente na *Web*. As características dos Diretórios serão descritas

---

[11] Aqui estamos nos referindo não à totalidade de recursos disponíveis na Internet, mas sim, primeiramente, às informações disponíveis em páginas na *Web*.

[12] É comum na contratação de serviço de acesso à Internet, que a empresa direcione o *Browser* do usuário para o seu *site* (do provedor). Discutiremos um caso à parte, quando o desenvolvedor do próprio *Browser* tem essa prática, como por exemplo, a Microsoft Inc. com seu navegador, o Internet Explorer.

[13] http://help.yahoo.com/help/br/dir/basics/basics-03.html e http://www.dmoz.org/about.html respectivamente, sendo que a segunda faz parte de um projeto – *Open Directory Projet* (ODP) da Netscape Corp. Detalharemos melhor esta proposta no capitulo 3.

com mais detalhes adiante. Entretanto, adiantaremos que uma das suas limitações mais marcantes está na estrutura hierárquica rígida e na forma de organização imposta às páginas, sacrificando muito a flexibilidade e colocando o usuário à mercê de uma organização arbitrária, na qual as informações e páginas estão categorizadas conforme os objetivos e interesses dos que mantêm estas ferramentas. Entendemos que esta característica, ao mesmo tempo em que facilita a localização de uma dada informação, causa limitação das possibilidades de busca. Afinal, as páginas estão sendo categorizadas por critérios definidos pelos organizadores destes serviços, que por sua vez estariam vinculados a seus próprios interesses ou de um grupo atuante de usuários da *Web*[14].

A terceira opção seria a de realizar uma busca automática pela *Web*, através de programas desenvolvidos com este objetivo. Nestes programas, através de uma interface, o usuário digita algumas "palavras-chave"[15] ou frases, possibilitando que o programa faça uma seleção na base de dados disponível e mostre os resultados desta pesquisa ao usuário. Estes são apresentados como uma página com *links*, que direcionam para páginas que estão associadas com as "palavras-chaves" fornecidas pelo usuário. Estas ferramentas são conhecidas como Mecanismos de Busca Automáticos, ou simplesmente Mecanismos de Busca, tendo como seus principais representantes atualmente: Google

---

[14] Referência às iniciativas de Diretórios colaborativos (DMOZ), que têm seus editores como voluntários.

[15] Este termo tem grande importância para as nossas argumentações futuras e será melhor definido e enquadrado em momento oportuno.

Search, Yahoo! Search e MSN Search[16]. Embora estes programas venham ganhando espaço frente aos demais acima citados, ainda estão longe de suplantá-los quanto à qualidade de seus resultados (MAZE, MOXLEY e SMITH, 1997, p. 10).

Das três possibilidades de pesquisa de informação na *Web* apresentadas, pode-se caracterizar como ferramentas formais de busca apenas o segundo e terceiro métodos, já que o primeiro método é apenas uma navegação livre pelo ciberespaço.

1.3. Estrutura de funcionamento das Ferramentas de Buscas

As ferramentas formais de busca podem ser categorizadas em dois grupos: Diretórios e Mecanismos de Busca Automáticos, embora seja importante ressaltar que ainda não há um consenso quanto à nominação desta separação, apesar dos métodos terem especificidades bem objetivas. Apesar de ser uma questão metodológica de abordagem do problema de busca na *Web* pelas empresas, a diferença entre as ferramentas está se tornando cada vez menos perceptível, mesmo para especialistas. Muitos técnicos especialistas e publicações feitas por empresas conceituadas na área[17], quando se referem a Mecanismos de Busca, englobam os dois métodos como se fossem apenas um, não fazendo esta diferenciação. Assim também, algumas empresas que inicialmente se caracterizavam por manter um serviço de Diretórios passam a lançar

---

16 Google Search – http://www.google.com; Yahoo! Search – http://www.yahoo.com e MSN Search – http://www.msn.com

17 Por exemplo, a comScore, uma provedora de informações e consultoria, que analisa a participação das empresas de "mecanismos de busca" no mercado norte-americano.

mão de Mecanismos de Busca Automatizados de parceiros no mercado, ou então desenvolvem o seu próprio. Por sua vez, outras que se firmaram no mercado como Mecanismos de Busca Automáticos passaram a utilizar os serviços de Diretórios para ampliar a capacidade e relevância da sua ferramenta de busca. São exemplos notórios desta prática: a Yahoo! Inc., que associou-se e/ou adquiriu diversas empresas de Mecanismos de Busca Automáticos para garantir a busca no idioma chinês (ARNOLD, 2005). e a Google Inc., que lançou mão dos resultados do Diretório do Projeto de Diretório Aberto (*Open Directory Project*) DMOZ em composição a seu Mecanismo de Busca Automático[18].

A não distinção dos dois métodos de busca mencionados (Diretório e Mecanismo de Busca Automático) decorre, possivelmente da tentativa dessas empresas de suprir limitações de cada uma das ferramentas. Assim, tendem a mesclar, algumas vezes com sucesso, características de ambos os métodos de organização e disponibilização das informações. Parece-nos, a princípio, que o usuário comum que lança mão dos serviços na *Web* através destas diferentes ferramentas não está a par das estruturas e interesses que interferem na apresentação das informações por ele encontradas na rede. Por exemplo: o Mecanismo de Busca Automático do Google pode ser escolhido pelo usuário tão somente pela sua popularidade, sendo que os resultados através dele obtidos podem parecer para este mesmo usuário muito similares aos obtidos através do Yahoo! que é primordialmente um Diretório. Assim, para este estudo, julgamos importante se fazer uma distinção e ressaltar as peculiaridades quanto às

---

[18] A Google faz essa afirmativa em sua página de ajuda aos seus serviços.

formas de inclusão e de manipulação dos dados de cada um dos sistemas e como são apresentados.

A seguir são detalhados os funcionamentos dos Diretórios e dos Mecanismos de Busca Automáticos.

1.3.1. Os Diretórios

O termo Diretório vem da forma como as informações, constituídas pelas páginas *Web*, são organizadas e apresentados ao usuário por esta ferramenta de busca. Trata-se de um "diretório" hierárquico, no qual os assuntos são organizados em categorias e subcategorias. As categorias, também chamadas de *Headings*, são dispostas em uma "árvore", de tal forma que o usuário possa fazer uma "navegação" direcionada através delas até o conjunto de *links* de páginas do assunto objeto de sua pesquisa. Na estrutura do Diretório, parte-se de uma categoria mais geral para uma mais específica.

Figura 1 - Página principal do DMoz[19]

A figura 1 mostra a página principal de uma proposta colaborativa que se utiliza da estrutura de Diretórios para organizar e apresentar as informações disponíveis na *Web*.

Procurando-se informações sobre "Tecnologia e Sociedade" utilizando um Diretório como o DMOZ, tem-se em sua página principal as categorias: *Arts, Business, Computers*, etc, onde é possível escolher a categoria *Science*. Seguindo este *link*

[19] http://www.dmoz.org

encontra-se neste nível as subcategorias: *Chats and Forums, Museums@*[20]*, News and Media , Science in Society*, etc. Neste ponto, seguindo pelos *links* tendo em vista obter a informação desejada e a similaridade com as subcategorias existentes, serão encontrados no final da busca os *links* para as páginas com assuntos relacionados ao que foi buscado. Este é uma das possibilidades de busca nos diretórios. Outra maneira de fazer busca em um diretório será descrita mais adiante.

Uma especificidade dos Diretórios é a presença de editores que atuam diretamente na inclusão e classificação das páginas. É a utilização do elemento humano na avaliação direta de página por página, com o objetivo de fazer o seu posicionamento em uma estrutura hierarquizada por assunto, que é uma das principais distinções entre esta ferramenta e os Mecanismos de Busca Automáticos. Nos primeiros Diretórios, havia apenas uma ordenação alfabética das páginas *Web* dentro de cada categoria. Atualmente, é possível visualizá-las também por "importância". No entanto, não são divulgados os critérios desta ordenação.

Para que uma página possa ser incluída neste catálogo de assuntos é necessário que o responsável a submeta[21] aos critérios dos editores deste Diretório. Ela será analisada, podendo ser aprovada ou rejeitada para fazer parte da base de dados daquela empresa de ferramentas de busca. Em uma segunda etapa, ela será indexada (ordenada)

---

20 O símbolo "@" no final do nome da subcategoria, especifica que esta pertence a outras categorias do Topo, desta forma, pode-se chegar a ela optando pela categoria *Arts*.

21 O procedimento de submissão requer a apresentação do endereço da página com a descrição de seu assunto ou conteúdo. É também necessário o cadastramento no *site* dos proprietários ou responsáveis pelo Diretório. É importante esclarecer que os procedimentos podem ter ligeiras diferenças, dependendo da empresa ou responsáveis pelo serviço de Diretório.

de acordo com os critérios utilizados pela empresa. Essa avaliação define tanto a classificação da página no ordenamento (*ranking*) daquele Diretório, como também a enquadra em uma categoria por assunto. Os critérios utilizados não são "amplamente divulgados", e "variam de acordo com os diferentes editores e empresas" (INTRONA E NISSENBAUM, 1998, p. 172). Há que se destacar também as interferências de toda ordem da variável humana a que fica submetido este ordenamento de páginas que será apresentado ao usuário, questão que foge ao escopo deste trabalho.

Através da navegação pelas categorias de um Diretório, há uma maior possibilidade e facilidade do usuário encontrar uma informação que procura, em um volume não muito extenso de informação e para os assuntos divididos ou classificados desta forma (YANG e LEE, 2003). No entanto, por ser um processo caracterizado pela intervenção humana direta com avaliação *site* a *site*, há uma limitação no número de *sites* disponíveis por esse método, diminuindo o dinamismo do sistema se comparado à velocidade e volume de informações processadas pelas ferramentas automatizadas. Sendo assim, a variação e as possibilidades de se encontrar assuntos e temas também se restringem, bem como a capacidade de indexar o crescente número de páginas que a cada dia são criadas e disponibilizadas na *Web*.

Quanto às limitações impostas pelo sistema de Diretório, cabe destacar a sua característica primeira, que é a necessidade da apresentação (submissão) formal pelo responsável pelo *site* aos critérios das empresas desta ferramenta de busca, para que este seja avaliado. Esta imposição associada ao crescimento muito rápido da *Web*, impossibilita a inclusão de muitas páginas no banco de dados das organizações que mantêm este tipo de ferramenta. Muitas limitações estão sendo superadas pelos Diretórios em funcionamento, tendo em vista que eles se tornaram ferramentas híbridas,

diversificando seus métodos de busca. Dentre estas diversificações, podemos destacar a que permite a busca por palavras na página principal da ferramenta. Por exemplo, a empresa Yahoo! Inc. apresenta em sua ferramenta de busca (Diretório) uma interface com o usuário, onde há um espaço em que pode-se introduzir palavras ou frases, como nos Mecanismos de Busca Automáticos[22]. Na Figura 2 pode-se observar, no destaque, a localização da área onde introduzir palavras-chaves para busca no Diretório. No entanto, esta é somente uma outra forma do usuário fazer uma busca **apenas** na base de dados do próprio Diretório. Como já comentado neste capítulo, com a aquisição de empresas e Mecanismos de Busca como o AltaVista, Inktomi e AllTheWeb[23] dentre outros pela Yahoo! Inc., esta passa a prestar também serviços de Mecanismo de Busca Automático na mesma interface de sua ferramenta (Diretório), apresentando, desta forma, uma interface única onde a solução de Busca Automática funciona juntamente com a estrutura de Diretório. Arnold (2004) critica a Yahoo! Inc. como sendo apenas um aglomerado de soluções de ferramentas de busca na *Web* compradas de outras empresas que tiveram algum sucesso nesta área. De qualquer forma, entendemos que independentemente da abordagem de cada empresa, esta é uma tentativa de minimizar as limitações de suas ferramentas de busca na *Web*.

---

22 http://help.yahoo.com/help/us/dir/basics/basics-03.html.

23 Das empresas Digital Equipments, Inktomi e FAST Serch&Transfer, respectivamente.

Figura 2 – Busca através de palavras-chave no Diretório do Yahoo! [24]

1.3.2. Os Mecanismos de Busca Automáticos

Os Mecanismos de Busca Automáticos têm como característica diferenciadora dos Diretórios a capacidade de busca automática de páginas *Web* para inclusão em suas bases de dados. Posteriormente, estas informações são disponibilizadas para pesquisa através de palavras ou frases apresentadas pelo usuário em sua consulta, Maze, Moxley e Smith (1997, p. 13) desmembram o funcionamento dos Mecanismos de Busca

[24] Página arquivada de 17 de outubro de 1996 (http://web.archive.org).

Automáticos em três partes básicas (vide esquema na Figura 3) - que são realizadas automaticamente por *softwares*: coleta, ordenação e apresentação:

- encontrar e pesquisar páginas na *Web* – esta etapa consiste na coleta de informações pela ferramenta de busca, sendo executada por programas comumente chamados de robôs (*robots*), que também são conhecidos por *spiders, crawlers, softbots,* etc (PINKERTON *apud* CHO, GARCIA-MOLINA e PAGE. 1998). A nomenclatura para estes programas mudam principalmente por razões comerciais e/ou na tentativa de individualizar alguma característica técnica ou comercial, como é o caso da Google Inc., que chama o seu robô de *Googlebot*. Aprofundaremos este tema mais à frente

- indexar as informações das páginas coletadas e possibilitar o cruzamento deste índice com as informações introduzidas pelo usuário na ferramenta de busca - nesta etapa os dados são organizados / ordenados em um repositório para serem apresentados, e também são utilizados como retro-alimentação dos robôs em suas coletas.

- prover uma interface para a consulta feita pelo usuário – nesta fase do processo, as páginas (ou *links* para as mesmas) são apresentadas ao usuário. Este aspecto será desenvolvido com mais detalhes no capítulo 3, no qual serão colocadas questões a respeito das relações comerciais das empresas que disponibilizam ferramentas de busca na *Web*.

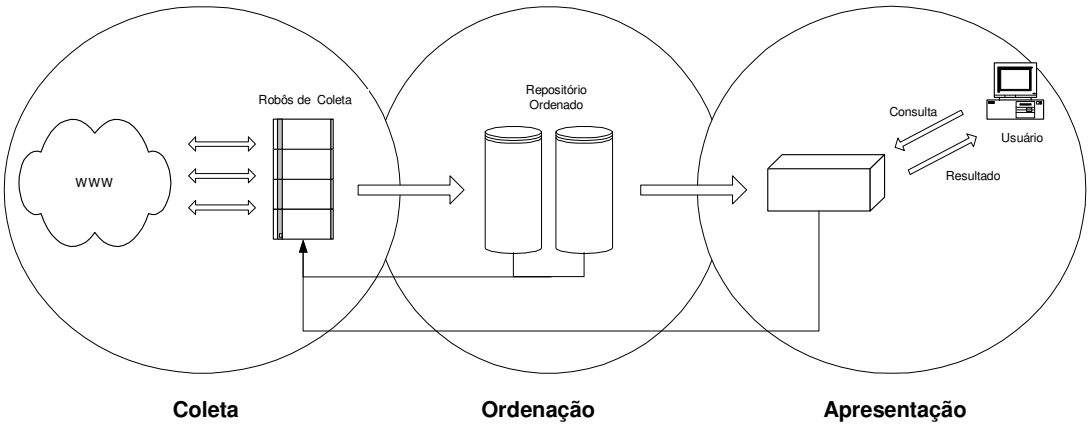

Figura 3 - Diagrama esquemático do funcionamento de um Mecanismo de Busca Automático padrão

1.3.2.1. Os Robôs dos Mecanismos de Busca

Os robôs são programas que fazem parte de um conceito mais amplo, os "agentes" (WOOLDRIDGE e JENNINGS, 1995), desenvolvidos para atuar de modo autônomo. O conceito de agente é controverso, não só pela utilização abrangente do termo, como pelo enfoque de implementação e aplicações do mesmo (NWANA, 1996, ARASU *et al*, 2001). Desta forma, abordaremos com mais detalhes apenas os robôs que estão associados aos Mecanismos de Busca Automáticos, não nos detendo nas questões da autonomia, independência, e aprendizado destes programas.

Os robôs inicialmente eram orientados apenas para vasculhar a *Web* em busca de páginas, como se fossem navegadores automáticos, similares a um usuário humano seguindo *links* para alcançar diferentes páginas, trazendo apenas algumas informações a respeito das que visitaram, como exemplo, o título da página e os *links* contidos nela. Posteriormente, na tentativa de tornar mais relevantes os resultados das consultas dos

usuários, uma quantidade cada vez maior de informação começou a ser coletada destas páginas. Esta atitude, entretanto, passa a acarretar o aumento do tráfego e, conseqüentemente, a redução da performance de acesso a um *site* no momento em que este está sendo pesquisado pelos robôs (KOSTER, 1995). Isto se tornou uma preocupação para os seus administradores (dos *site* e dos robôs) e que permanece ainda hoje, a ponto de existirem procedimentos e práticas para impedir ou controlar o acesso destes robôs ao vasculhem os *sites* alvos[25].

O interesse das empresas e/ou pessoas de serem indexadas por um Mecanismo de Busca Automático ou Diretório é notório. No entanto, para algumas empresas não há interesse neste volume adicional de tráfego que pode comprometer o serviço prestado para os seus usuários legítimos e possíveis consumidores. Ainda mais havendo a possibilidade de não existir nenhum acordo de parceria na utilização das informações disponíveis nestes *sites*, logo também poucas vantagens diretas para o *site* alvo. Mesmo instituições públicas como universidades, que possuem recursos financeiros limitados e são alvos bastante freqüentes dos Mecanismos de Buscas, podem não ter interesse em que seu tráfego e/ou recursos sejam consumidos por ferramentas de empresas com objetivos apenas comerciais.

---

[25] Os robôs que seguem os padrões definidos no robots.txt protocol, obedecem a regras de exclusão de páginas definidas pelos administradores dos *sites*, deixando para estes administradores a incumbência de permitir ou não a varredura de seus *sites* pelos robôs dos mecanismos busca na Web. O protocolo é apenas uma recomendação, não estando os robôs impedidos de acessar as páginas de um *site*. Isso ocorre pelo fato dos robôs simularem o comportamento de um usuário legítimo, sendo muito difícil um possível bloqueio caso alguma empresa não implemente este protocolo em seus robôs de coleta.

Os robôs não somente vasculham a *Web* a procura de novos *sites* ou páginas para incluí-los no banco de dados do mecanismo, como também os revisitam para atualizar as mudanças que possam ter ocorrido desde a última checagem. A freqüência desta atualização varia de um Mecanismo de Busca para outro, dependendo de critérios definidos por cada empresa. Algumas empresas de Mecanismos de Busca mantém um *cache* destas páginas[26], isto é, uma copia destas paginas disponíveis em seus próprios servidores, para que não ocorram inconsistências no resultado da busca, no caso do usuário tentar acessar um *link* desatualizado ou não disponível naquele momento. Desta forma, o mesmo pode optar por acessar um *link* que lhe apresentará um "instantâneo" da página indexada, isto é, como esta se apresentava quando a mesma foi capturada pelo robô. Ao longo dos últimos anos, este recurso, bem como muitos outros foram e vêem sendo desenvolvidos, acoplados, e oferecidos como novas funcionalidades e diferenciais de uma empresa para outra, objetivando facilitar a pesquisa do usuário e, por sua vez, garantir a fidelidade deste como cliente neste novo mercado de serviços.

Na base de dados dos Mecanismos de Buscas Automáticos ficam armazenadas uma grande quantidade de informação das páginas acessadas e, em alguns casos, de todo o *site*, quando este é totalmente indexado. Entretanto, um estudo feito com os principais mecanismos de busca norte-americanos - especificamente Google, Altavista e AllTheWeb - por Vaughan e Thelwall (2004), demonstra que o *software* de busca, ao acessar um determinado *site*, não garante que este seja totalmente indexado, como induz o discurso apresentado pelas empresas que disponibilizam estas ferramentas.

---

[26] informação divulgada no *site* do Google em http://www.google.com.br/webmasters/remove.html, onde primeiramente disponibilizou este tipo de serviço em sua ferramenta de busca.

Comparando a atuação das ferramentas de busca quando estas indexam *sites* nos EUA e na Ásia – especificamente China, Taiwan e Singapura - verificou-se que as buscas feitas nos EUA são mais profundas, ou seja, indexava mais páginas por *site*, em comparação aos *sites* dos referidos paises na Ásia, cobrindo entre 80% a 87% das páginas dos *sites* nos primeiros e de 4% a 75% nos segundos. Esta "característica técnica" tem repercussões nos resultados das buscas apresentadas ao usuário. Este aspecto é de grande interesse para as discussões que faremos no decorrer deste trabalho. Sendo a Google Inc. uma empresa norte-americana, fica caracterizado o seu interesse preferencial pelo mercado norte-americano, no que dá mais ênfase e direcionamento a sua ferramenta de busca.

O controle das empresas comerciais de mecanismo de busca feito de forma centralizada pode gerar distorções como as assinaladas por Jeanneney (2006), quando discute o sistema de digitalização de obras literária pela Google Inc. Jeanneney, que é Diretor da Biblioteca Nacional Francesa, levanta a hipótese que a empresa privilegiará obras de interesse norte-americano.

1.3.3 Outras Ferramentas de Busca

As empresas que oferecem serviços de busca na *Web* têm suas ferramentas de busca apresentando resultados que podem ser bastante diferentes para uma mesma pesquisa feita por um usuário, este comportamento ocorre principalmente por estas utilizarem critérios bem diversificados nestes serviços (MAZE, MOXLEY e SMITH. 1997). Apostando nestas diferenças, proclamam oferecerem os "melhores" e mais

relevantes resultados nas buscas apresentadas para a consulta feita por seus usuários em suas ferramentas.

Com este mesmo argumento, surgem empresas e grupos apresentando novas ferramentas de busca na *Web*, que se distinguem dos dois tipos já citados - os Diretórios e os Mecanismos de Busca Automáticos. Entre as quais podemos citar os Metabuscadores, cujo modelo de funcionamento é apresentado esquematicamente na Figura 4.

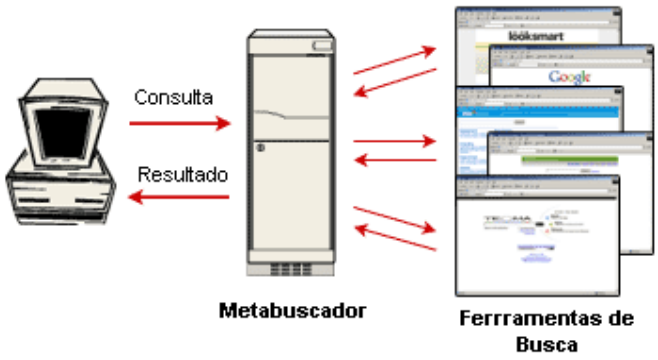

Figura 4 – Esquema de funcionamento de um Metabuscador.

O diferencial relativo a esta ferramenta está em que a mesma pode unir vários critérios de busca e ordenação de diversas ferramentas de busca. Desta forma, com apenas uma única consulta o usuário teria acesso à composição de diversos resultados de várias ferramentas, bem como da reorganização destes resultados, que só poderiam ser obtidos se ele executasse a mesma pesquisa em diversas ferramentas de busca individualmente, o que acarretaria em dificuldades adicionais como por exemplo, a de identificar e localizar na *Web* cada uma destas ferramentas e como saber se as mesmas

coletam as informações pesquisadas [27] e, possivelmente aprender a manusear as diferentes interfaces de cada uma (DREILINGER e HOWE, 1997). Com os Metabuscadores, pode-se, através de uma única ferramenta e interface acessar o resultado de todas ao mesmo tempo e em um único local.

Os Metabuscadores também podem, ao organizar resultados obtidos de várias ferramentas de busca, eliminar possíveis elementos duplicados. Além de fazer uma junção destes resultados, alguns Metabuscadores também os apresentam em categorias, simulando uma divisão por assunto como em um Diretório. No entanto, a utilização estes recursos adicionais acarretam em lentidão na apresentação da página com os resultados da consulta e nem sempre a categorização é assertiva e rigorosa, por ser feita automaticamente via *software*. Tem-se como exemplos de Metabuscadores disponíveis na *Web* o Dogpile, o Vivisimo e o Mamma[28].

Em um primeiro momento, os Metabuscadores parecem ser uma excelente alternativa para a busca na *Web*. Porém, o caminho tomado por algumas das empresas que detêm os principais e mais populares ferramentas de busca deste tipo resultou em uma opção extremamente polêmica. Os resultados obtidos junto às ferramentas de busca na *Web* eram associados a *links* patrocinados, ou seja, uma empresa poderia "comprar" uma inclusão ou um melhor posicionamento na página de resultados[29] da pesquisa feita

---

[27] Veremos adiante, o caso de ferramentas de busca para assuntos específicos.

[28] http://www.dogpile.com, http://www.vivisimo.com, http://www.mamma.com, respectivamente.

[29] A divulgação de uma pesquisa mostrando o percentual de resultados mostrados na página de busca que são propagandas pagas, nos principais metabuscadores disponíveis na Web está em (http://searchenginewatch.com/searchday/article.php/3109441).

por um usuário desta ferramenta. Muitas destas inclusões não eram explicitadas como tal para o usuário do Metabuscador, podendo passar como resultados legítimos obtidos nas ferramentas de buscas pesquisadas. Essa atitude desagradou tanto aos usuários, quanto às empresas e aos grupos que detinham as ferramentas de busca envolvidas. Possivelmente, pelo fato desses últimos não participarem com desejavam de eventuais vantagens comerciais relacionadas, algumas dessas empresas passaram a proibir a utilização dos resultados obtidos por suas ferramentas nestas circunstâncias.

Quando citamos os Mecanismos de Busca Automáticos e Diretórios, estamos nos referindo a categorias de ferramentas gerais, com objetivos de busca gerais. No entanto, há esforços também para o desenvolvimento de ferramentas de busca especializadas, que são orientadas a um único assunto ou temas correlatos. Estas ferramentas são pouco conhecidas, sendo empreendimentos individuais e/ou restritos a usos comerciais, ou ainda, com objetivos acadêmicos.Neste campo podemos encontrar, tanto sistemas com características de Mecanismos de Busca Automáticos como Diretórios. Quando estas ferramentas têm características de um Mecanismo de Busca Automático, a prospecção de *sites* permitirá melhores resultados que os seus similares generalistas. Quando, por outro lado, são Diretórios com assuntos específicos, acabam se tornando referência no tema, sendo utilizadas por Mecanismos de Busca generalistas para o início de sua prospecção de dados para um determinado tema[30]. Uma grande dificuldade é encontrar estes Diretórios específicos, que na sua maioria não são comerciais, mas empreendimento de professores ou pesquisadores de universidades ou

[30] Segundo o Prof° Ricardo Arantes (COPPE/UFRJ), que é responsável por um diretório *Web* especializado em referencias bibliográficas de assuntos e materiais acadêmicos.

organizações não comerciais. Estes por terem dificuldades de financiamento ou patrocínio, têm necessidade de mudar de locais e endereços freqüentemente.

Uma outra limitação para essas ferramentas é que, por terem uma orientação muito especifica, acabam por servir a um grupo muito restrito de usuários, tornando-se de pouca utilidade para a maioria dos usuários.

1.4. Os critérios para coleta, ordenação e apresentação

A descrição dos critérios de funcionamento de uma ferramenta de busca é de grande relevância, principalmente no que se refere à distinção do que é declarado oficialmente e a que é nos revelada na prática. Seria possível a ausência de critérios nos mecanismos de busca vir a se configurar em um critério? Talvez o não critério seja um critério. Afastadas as discussões metafísicas, quando uma empresa responsável por uma ferramenta de busca divulga que utiliza certos critérios específicos no desenvolvimento de seu *software*, nada nos garante que o mesmo se comportará de acordo com o que é divulgado.

Há na literatura especializada (ARASU *et al*, 2001, CHO, GARCIA-MOLINA e PAGE, 1998, INTRONA E NISSENBAUM, 1998), a utilização do termo "métrica" para determinar a relevância de uma página, possivelmente para dar uma conotação técnica à implementação de critérios para o funcionamento das ferramentas de busca na *Web*. As métricas definem como são orientados os programas para sua maior otimização na coleta e apresentação de páginas relevantes ao usuário da consulta.

Como descrito na seção 1.3.2, o processo de busca na *Web* é realizado em três etapas, podendo existir critérios distintos para cada uma delas. Pretendemos descrever alguns destes critérios ou categorias de critérios, que consideramos relevantes para o esclarecimento de nosso objetivo quanto ao funcionamento das ferramentas de busca.

O funcionamento dos mecanismos de busca pode se basear em centenas de critérios para determinar a inclusão e relevância das informações apresentadas. Estes critérios podem ser divididos em três categorias básicas: popularidade; forma e palavras; e conteúdo (BRANDT, 2002). No entanto, esta divisão assim como o próprio termo critérios, já comentado neste capítulo, são usados de maneiras diferentes por diversos autores. Portanto, esta divisão não é unânime na literatura da área.

A obscuridade ou omissão dos critérios, está na forma como estes são implementados e combinados entre si para definir o funcionamento da ferramenta. Este funcionamento está normalmente ligado ao objetivo especifico da empresa que o implementa. Podemos encontrar na literatura a descrição de uma grande variedade de critérios que são utilizados nas ferramentas de busca. Temos, por exemplo, os que objetivam suprir demandas sazonais[31]. Outros objetivos visam impedir que haja

[31] Como um dos objetivos das empresas de ferramenta de busca é atender o maior número de pessoas possível, pode ser que uma informação de pouca "relevância" para a maioria, em uma ocasião se torne de grande interesse, como o caso da expressão "World Trade Center" ou WTC, nos dias que se sucederam ao 11 de setembro de 2001. Certamente as ferramentas de busca tiveram que ser redirecionadas para esta demanda de informação que, no entanto, é sazonal e necessária para aquele momento.

manipulação por parte de outras empresas ou grupos (*Spammer*) [32], que desejam figurar nas listas das buscas.

O levantamento de critérios usados pelas diversas ferramentas de busca na *Web* é o objeto de estudos e pesquisas dos Otimizadores para Mecanismos de Busca[33] (*Search Engine Optimizer*, SEO), que oferecem seus serviços para empresas ou pessoas que desejam garantir tanto a inclusão como um melhor posicionamento nos resultados apresentados por estas ferramentas. No entanto, as empresas que comercializam métodos para o posicionamento ou reposicionamento de páginas na apresentação das consultas pelas principais ferramentas de busca, oficialmente não garantem a efetividade dos resultados deste serviço. Tanto as empresas de ferramentas de busca na *Web* , bem como as empresas de SEO, negam ter conexões entre suas atividades. Voltaremos a este ponto no capítulo 3, quando discutiremos em mais detalhes as questões comerciais que envolvem as ferramentas de busca na *Web*.

1.4.1. Popularidade

No caso das páginas *Web*, o critério de popularidade [34] está baseado na quantidade de vezes que esta é referenciada através de *links* de outras páginas na *Web*.

---

[32] Podemos verificar por exemplo que a Google Inc.e o Yahoo! Inc. justificam em suas páginas de ajuda a necessidade de evitar a ação dos *spammers* através da adoção de determinados critérios.

[33] Em http://www.google.com/webmasters/seo.html a Google Inc. também se refere aos SEOs.

[34] *Link popularity* – Parece-nos que este termo em inglês ainda não encontrou (como muitos outros da área) uma similaridade ou estabilidade na literatura em português.

O grau de relevância, neste caso popularidade, de uma página **P** é calculado pelo número de *links* que são direcionados para esta página. É o número de *links* de entrada ou *Backlinks* da página **P** (INTRONA e NISSENBAUM *apud* CHO *et al*. 1998, pág. 172), que determina o posicionamento desta página, isto é - a sua importância ou destaque em relação a outras páginas. O termo *Inlink* também é utilizado, neste caso, como referência de relacionamento entre as páginas (BJÖRNEBORN e INGWERSEN, 2004).

A popularidade de uma página usando o critério de relevância é empregado em um extenso número de mecanismos de busca (BRANT. 2002). O sistema de algoritmos conhecido como *PageRank*[35], base do mais popular mecanismo de busca[36] atualmente - o Google - tem como característica o emprego deste.

Passamos a seguir a fazer uma breve descrição do *PageRank*. Como muitas das outras funcionalidades dos demais mecanismos de busca, os detalhes também não são explicitados pelos que os aplicam. No entanto, muitas análises feitas a respeito do seu funcionamento levam em consideração um artigo[37] escrito pelos fundadores do Google (Brin e Page, 1998), quando estes eram alunos da Universidade de Stanford, onde

---

35 Segundo o Wikipedia (http://wikipedia.org), o termo *PageRank* é um anacronismo ligado ao nome de um de seus desenvolvedores – Lawrence (Larry) Page, mas isto é apenas uma especulação.

36 segundo a comScore (http://www.comscore.com) resultados de julho de 2005.

37 *The anatomy of a large-scale hypertextual Web search engine*, artigo apresentado na Sétima Conferência da *World Wide Web* (WWW7) em Brisbane, Australia.

desenvolveram a base do algoritmo do *PageRank*. O registro de patente do *PageRank*[38], também nos dá uma idéia de seu funcionamento*:*

> "A method assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database. The rank assigned to a document is calculated from the ranks of documents citing it. In addition, the rank of a document is calculated from a constant representing the probability that a browser through the database will randomly jump to the document. The method is particularly useful in enhancing the performance of search engine results for hypermedia databases, such as the world wide web, whose documents have a large variation in quality". (United States Patent and Trademark Office - USPTO, 1998)

O *PageRank* é considerado um método de medição da "importância" de uma página. A principal característica deste método está em desenvolver o cálculo de um valor para cada uma das páginas da *Web*, o qual é dado pela relação de citação entre as páginas na *Web.* Este índice varia entre 0 e 10, sendo que a página que tiver *PageRank* igual a 0 é a de menor valor e a que possuir *PageRank* igual a 10 é a de maior valor[39] (RIDINGS e SHISHIGIN, 2002). Assim, quanto maior for o valor dado a uma página, tanto maior será sua relevância comparada a outras páginas.

Desta forma, se uma página A tem um *link* para uma página B, o método considera que a página A está dizendo que a página B é uma página importante, logo ela ganha um "voto". Até este ponto, continua igual ao critério de popularidade, já usado

---

[38] Publicada no *site* governamental US Patent & Trademark Office (http://www.uspto.gov), com o registro pertencente à Universidade de Stanford e Lawrence Page, como o inventor.

[39] Estes são apenas os valores de representação do *PageRank* que é disponibilizado pela Google Inc. através de sua barra de ferramentas. Os valores reais do *PageRank* nunca foram apresentados (RIDINGS e SHISHIGIN. 2002, pág. 5)

por muitos mecanismos de busca até então. O diferencial está em que um voto dado de uma página de destaque neste critério na *Web* tem mais valor que um outro de uma página de menor importância. Ou seja, um *link* vindo de uma página C, pouco referenciada, para a página B tem peso menor do que se o *link* vem da Página A, se esta for muito referenciada. Assim, os votos ou citações não têm o mesmo peso. Logo, uma página com um *PageRank* baixo pode ter seu valor aumentado significativamente se for referenciada por outra com um *PageRank* de alto valor mesmo que haja poucas referencias de outras páginas para ela.

Portanto, o destaque da página não será medido apenas pelo número de referências a ela, mas também pela importância, isto é, peso de cada referência. Se, por um lado, esta ponderação pode parecer razoável, por outro, a indeterminação envolvida na medição abre espaço para manipulação interna.

O *PageRank* tornou-se um paradigma de utilização para os mecanismos de busca atuais (EIRON, MCCURLEY e TOMLIN, 2004). Apesar de algumas diferenças de implementação por parte de cada empresa, todas se utilizam deste critério para definir, em algum momento, a ordenação ou coleta de páginas pelo seu mecanismo de busca.

1.4.2. Características da página (Forma e palavras)

Forma e palavras fazem parte da categoria de critérios que levam em consideração as características intrínsecas da página (BRANDT, 2002). Incluem especificidades ligadas ao tamanho e/ou tipo das fontes utilizadas na página, freqüência das palavras no texto, proximidades das palavras, nome do documento relativo à página,

nome do diretório, nome do domínio ou URL. No passado, este critério foi o mais difundido e utilizado nas ferramentas de busca.

Figura 5 – Características de uma página *Web*[40]

Como descrito anteriormente, para a codificação de uma página *Web* é usado o HTML ou derivações deste, onde é possível, através da análise das *tags*, a verificação de características tais como o tipo e/ou tamanho da fonte que estão sendo usadas, bem

---

40 A página *Web* deste exemplo não está *online*.

como das suas cores. Por exemplo, uma página simples como a apresentada na figura 5 teria a seguinte codificação em HTML[41]:

```
<HTML>
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.0//EN" "http://www.w3.org/TR/html4/strict.dtd">
<HEAD>
<TITLE>Exemplo para seguir o Site do HCTE </TITLE>
<META NAME= "HCTE" CONTENT="Donizeti Batista">
</HEAD>
<BODY>
<H1>Cabeçalho da Página Web</H1>
<P><H3>Esta já é a segunda linha do texto!</H3></P>
<P><a href="http://www.ccmn.ufrj.br/hcte/index.htm" Este link aponta para a página do <B>HCTE</B>.</a></P>
</BODY>
</HTML>
```

O texto entre as tags **<H1> ...</H1>** é apresentado com maior destaque do que o que está entre as marcações **<H3> ...</H3>. D**esta forma os mecanismos de busca podem ser programados para dar mais importância às palavras que estão na frase com maior destaque e colocá-las como representantes do tema a que se refere esta página, bem como o texto entre **<TITLE> ...</ TITLE>** é utilizado da mesma forma por ser o título da página. Estas técnicas baseadas neste critério permitem que, através de uma avaliação simples e automática, seja possível obter informações sobre o assunto tratado na página. O exemplo da figura 5 mostra alguns problemas associados a esta metodologia, pois se esta página fosse indexada por um mecanismo de busca, apesar de o título fazer referencia ao "Programa de História das Ciências, das Técnicas e

---

41 Nosso objetivo aqui de não é ensinar codificação HTML, mas sim, apresentar como os robôs dos mecanismos de busca podem obter informações de uma página *Web* codificada desta forma. Assim apontaremos apenas as codificações que forem relevantes ao assunto que queremos abordar. Esta versão de código é para o HTML 4.0 (MARCONDES, 1998).

Epistemologia" ela não diz nada ou quase nada a respeito deste assunto. No entanto, poderiam ser associados a estes critérios, outros tais como a existência e freqüência de palavras ligadas ao tema, a sua proximidade no texto, as fontes utilizadas nas palavras ou se estas estão em negrito, dentre outros que permitiriam uma classificação ou análise mais adequada.

Levando em consideração que a *Web* é um espaço virtual multimídia, onde as páginas podem conter, além de texto, também imagem e sons associados a ela, uma página que contenha muitas fotos pode ser considerada de relevância para um usuário que pesquise sobre *design* gráfico. Logo, pode não ser apenas o texto um fator de relevância para uma página. Apesar disso, em alguns casos, avaliando apenas a forma e os elementos existentes em uma página é possível identificar e calcular a sua relevância. Este foi um dos primeiros e principais critérios utilizados pelos mecanismos de busca automáticos e que ainda são largamente utilizados. É o que veremos mais adiante no capítulo 3, quando discutiremos as questões comerciais das empresas de ferramentas de busca na *Web* e a utilização das palavras chave como instrumento de propaganda.

1.4.3. Análise de conteúdo

No item anterior foram descritas as características de como um mecanismo de busca poderia classificar uma página através da forma como esta se apresenta. Esta analise é extremamente precária. Por exemplo, uma dada página onde a freqüência de aparição da palavra "Leão" é alta, não é uma página de relevância para um usuário que estivesse pesquisando e interessado em informações sobre o animal africano. A referida página poderia ter informações do jogador e hoje técnico de futebol Emerson "Leão",

ou mesmo do mate "Leão" [42]. Assim, como resultado da pesquisa poderemos obter todas estas páginas que dão um significado diferente ao termo "Leão". Portanto, pode-se perceber, através deste exemplo, que entre o que o usuário deseja encontrar e o que lhe é oferecido, pode haver uma distância muito grande. Neste momento, já existem alguns mecanismos de busca que levam em consideração a questão do conteúdo semântico. Por exemplo, o Exploora[43], em uma consulta pela palavra "manga", permite ao usuário desmembrá-la em resultados de vestuário ou fruta. Estas iniciativas ainda são tímidas e necessitam de aprimoramentos. Há um movimento liderado pela W3C, através de seu fundador, Tim Berners-Lee, para o desenvolvimento de uma *Web* semântica onde fossem utilizados padrões através de *tags* do HTML ou suas derivações e ampliações (BERNERS-LEE, HENDLER e LASSILA, 2001; SOUZA e ALVARENGA, 2004). No entanto, o sucesso deste empreendimento dependerá da colaboração de empresas e organizações participantes deste sistema e nada garante que esta colaboração esteja de acordo com seus interesses comerciais.

---

[42] Um dado empírico sem uma metodologia aprimorada mostrou que o "*Top Spot*" do Google para a palavra "Leão" é a página do jornalista Leão Lobo(resultado de 21/12/2005).

[43] http://www.exploora.com.br.

Capítulo 2

MECANISMO DE BUSCA DO GOOGLE - LIMITAÇÕES E PROBLEMAS

Neste capítulo serão apresentadas algumas características ligadas a limitações e possíveis problemas do mecanismo de busca mais popular utilizado na *Web* hoje: o Google, que no mercado dos EUA é a ferramenta empregada em mais de 49% de todas as buscas feitas na Internet[44]. Embora tenham sido analisadas características das ferramentas de busca na *Web* de uma forma geral no capítulo anterior, serão retomadas aqui com mais detalhes aquelas utilizadas pela empresa Google Inc., tendo em vista aprofundar a discussão sobre as falhas e/ou limitações do serviço oferecido e questionar o discurso de eficiência veiculado pelos representantes desta empresa. Cabe destacar que a preocupação central deste trabalho se refere menos ao problema que surge quando um usuário, ao lançar mão da ferramenta de busca na *Web,* não consegue um resultado que considere satisfatório, e sim à situação em que este usuário encontra resultados que, a princípio, lhe parecem pertinentes.

Os resultados obtidos e apresentados ao usuário da ferramenta, com raras exceções, são filtrados em função de diversas "limitações". Serão destacadas as de ordem técnica, as surgidas para atender a interesses econômicos das empresas e as que são frutos da associação de ambas. Serão discutidas algumas características dos recursos técnicos utilizados e os discursos dos representantes da Google Inc, considerados de

---

[44] Fonte: Nielsen/NetRatings MegaView Search, NetRatings Inc. (http://www.comscore.com), relatório de julho de 2006. Em um total analisado aproximado de seis bilhões de consultas às ferramentas de busca na *Web* no mês de julho de 2006.

caráter oficial. Serão, ainda, trazidos dados de uma literatura independente que discute esta questão, visando confrontar os posicionamentos e apontar algumas contradições no discurso oficial apresentado pela Google Inc.

2.1. Limitações no acesso às informações na *Web*:

Inicialmente, serão destacadas algumas limitações de caráter geral com relação à obtenção de informações na *Web.* A divisão em tópicos, apresentada abaixo, objetiva atender, especialmente, a necessidades de ordem metodológica, Existe uma grande dependência entre os fatores limitadores destacados:

- Das características técnicas da ferramenta: limitação devido ao distanciamento entre a quantidade de informações disponíveis na *Web* e a capacidade física de *hardware* e o design do *software* dos Mecanismos de Busca de catalogá-las e apresentá-las ao usuário de forma acessível. Com o crescimento explosivo da Internet e, principalmente, da *Web*, os mecanismos de busca que indexavam até 95% das 19 milhões de páginas existentes em 1996 (CHU e ROSENTHAL, 1996), não indexavam mais de 42%, das 800 milhões de páginas disponíveis na *Web* em 1999, segundo estudos estatísticos feitos por Lawrence e Gilles (INTRONA e NISSENBAUM, 2000). Para se ter uma idéia da dificuldade de catalogar estas informações e disponibilizá-la, cabe acrescentar que em 2005 o número de páginas acessíveis na *Web* ultrapassou a cifra de 9 bilhões.

- Do usuário: desconhecimento das técnicas de funcionamento dos Mecanismos de Busca e do universo de informações disponível. O usuário,

normalmente, tem poucos dados sobre o tema que está buscando neste espaço virtual e não tem informações suficientes sobre o funcionamento da ferramenta. Isto faz com que ele não seja capaz de reconhecer o fato da busca frequentemente se distanciar dos seus objetivos, obtendo resultados parciais, resultados equivocados ou nenhum resultado.

- Da influência dos interesses econômicos e empresariais frutos das relações de parcerias comerciais comuns neste setor: os interesses comerciais da empresa proprietária da ferramenta ou seus parceiros e instituições associadas podem estar em contradição com os interesses dos usuários. Consideramos que esta contradição é um fator limitante da ferramenta e é a questão central que desejamos abordar neste trabalho. Um aprofundamento neste ponto será feito no próximo capítulo.

Apesar da amplitude das limitações apontadas acima, a Google Inc. busca atribuir ao usuário as responsabilidades pelos problemas enfrentados na localização de informações através de sua ferramenta. Por exemplo, o *Google Gulp*[45] (ver Figura 6), foi uma bebida fictícia criada pela Google Inc., que transfere ao usuário as limitações de seu sistema de busca, utilizando-se de ironias para as frustrações enfrentadas pelos usuários. Desta forma, responsabiliza-se o próprio usuário quando este se defronta com uma pesquisa mal sucedida.

---

[45] Lançado do dia 1º de abril de 2005 em http://www.google.com/googlegulp/, continua disponível ainda hoje (20/02/2007).

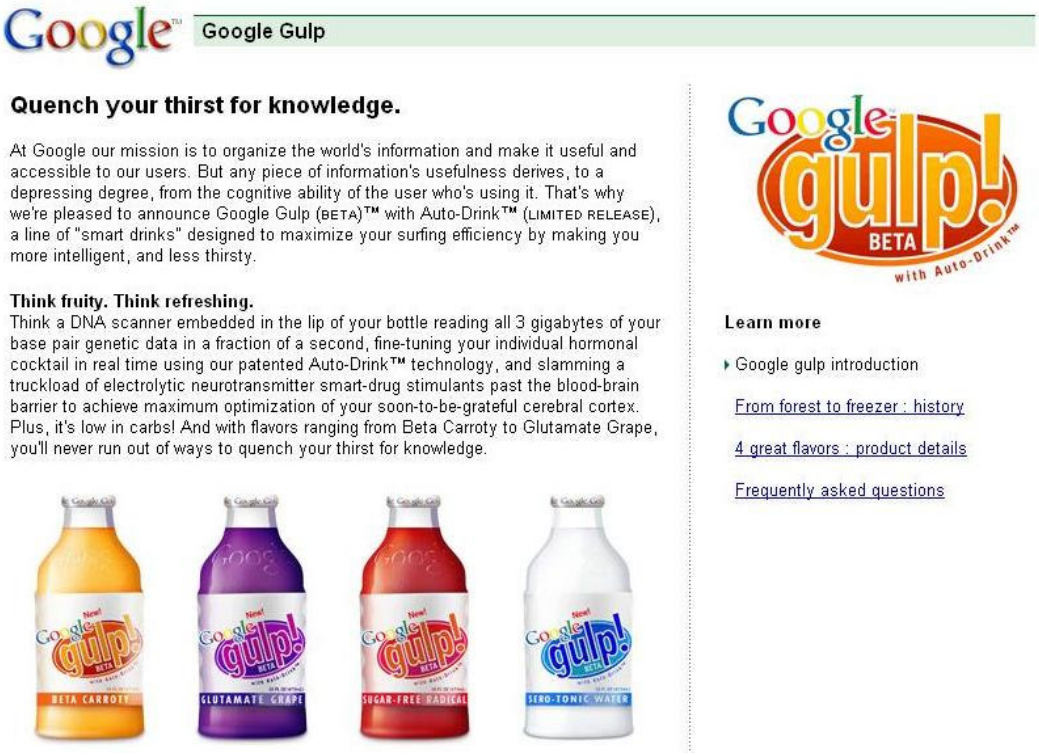

Figura 6 - A página com a propaganda da Google Gulp!

Apesar de concordarmos com Carvalho (2000) em seu livro "Datamining", no qual descreve que qualquer sistema de *data mining*[46], em última instância, depende do fator humano para seu êxito, pode-se afirmar que esta associação sugerida pela Google Inc. ignora o fato de que as limitações das pesquisas obtidas na *Web* não são apenas fruto das inabilidades técnicas (ou cognitivas, segundo a página da Google Gulp!) do usuário e sim de um conjunto de fatores como os apontados anteriormente.

---

46 Processo de extração automática de informação ou conhecimento de uma grande base de dados ou conjuntos de dados.

Cabe salientar que, no caso das ferramentas de busca na *Web,* a dificuldade se acentua tendo em vista o seu crescimento exponencial e que as informações neste espaço virtual estão disponibilizadas sem uma organização ou padrão pré-definido.

Devido à maneira como a Internet foi concebida e como funciona atualmente, a regulamentação mantém-se em níveis mínimos. Assim, as informações são disponibilizadas na *Web* sem que o autor tenha que fazer declarações sobre o uso, finalidade ou conteúdo do que está disponibilizando, nem tampouco apresentar palavras-chaves ou enquadrá-las em um índice. Evidentemente, este enquadramento poderia ser questionado em função da redução da informação a um saber enciclopédico. Corre-se também o risco de que a entidade responsável por esta regulamentação não representar de forma democrática os interesses dos diversos setores envolvidos. Entretanto, não há duvidas que facilitaria muito a organização.

O enquadramento acima referido está, no entanto, muito longe da situação atual em que a governança da Internet fica à mercê de interesses das leis de mercado e das forças econômicas dominantes.

Um dos aspectos da organização das empresas que disponibilizam ferramentas de busca na *Web* e que pode gerar limitações nos resultados obtidos é o seu comprometimento empresarial e publicitário. Este é um dos pontos que defendemos e que será tratado no Capítulo 3.

Os fundadores da Google Inc. questionam, em artigo acadêmico, o comprometimento comercial de outras empresas de mecanismo de busca, afirmando que os:

> "motores de busca migraram do campo acadêmico para o comercial. Até agora, a maior parte do desenvolvimento de

> motores de busca foi realizado em empresas, com pouca publicação de informações técnicas. Isso faz com que a tecnologia de máquinas de busca permaneça em grande medida uma arte obscura e seja orientada à publicidade." (PAGE E BRIN, 1997)[47]

No entanto, eles contrariaram suas próprias afirmações, como destaca o trabalho de Diaz-Isenrath (2005), quando comenta a "Carta dos fundadores da empresa aos futuros investidores" (*Letter from the founders:'an owner´s manual for Google shareholders*):

> "[...] Tudo parece como se tratasse de uma Enciclopédia-Mundo, que, no entanto, deixaria de lado uma parte da ambição universalista dos enciclopedistas. O conhecimento sobre o algoritmo e toda uma série de mecanismos e desenvolvimentos técnicos associados são, agora, zelosamente guardados por uma corporação.". (DIAZ-ISENRATH ,2005, pág. 114)

Embora Diaz-Isenrath faça um enquadramento da questão em uma linha sócio-técnica, que não é a abordagem utilizada neste trabalho, pretendemos incorporar seu questionamento e, se possível, ampliá-lo e aprofundá-lo. Deve-se indagar os limites da eficiência das ferramentas de busca na *Web* tendo em vista a influência da intrincada rede de interesses econômicos e com os quais as empresas estão envolvidas. Cabe examinar de que forma estas influências interferem nas escolhas dos padrões técnicos das ferramentas e nos resultados oferecidos aos usuários.

---

[47] Texto original traduzido por DIAZ-ISENRATH (2005).

2.2. O *Google Bomb*

2.2.1. O texto-âncora e o *Google Bomb*

Para abordar a questão do *Google Bomb* primeiramente será ralizada uma breve descrição ou definição do que é um *hiperlink* ou *link*, como é mais comumente chamado. Definimos o termo *link* como sendo uma referência dentro de uma página a uma outra página *Web* ou a uma parte desta (no caso de uma referência a uma determinada parte do próprio documento), ou a um *Web site*. É importante, neste momento, que não se confunda o endereço da localização de um *Web site* ou URL[48] com o *link* que o referencia. O segundo é uma referencia dentro de uma página, enquanto o primeiro é a sua localização na estrutura de nomes da Internet.

Os *links* de uma página *Web,* na maioria das vezes, não são mostrados explicitamente. Eles são apresentados como um texto explicativo do *link* que é denominado "texto ancora" (vide exemplo na Figura 7).

[48] *Relative Uniform Resource Locators* (URL), houve uma atualização nos padrões utilizados na Internet pela RFC2396, passando a utilizar a designação URI (Uniform Resource Identifiers) (BERNERS-LEE, FIELDING e MASINTER, 1998, p. 1), para referenciar os recursos na Internet (no caso, páginas Web, mas poderia ser um arquivo de vídeo, ou uma fotografia). No entanto, neste trabalho continuaremos a utilizar a denominação da sigla anterior, por sua tradição na literatura.

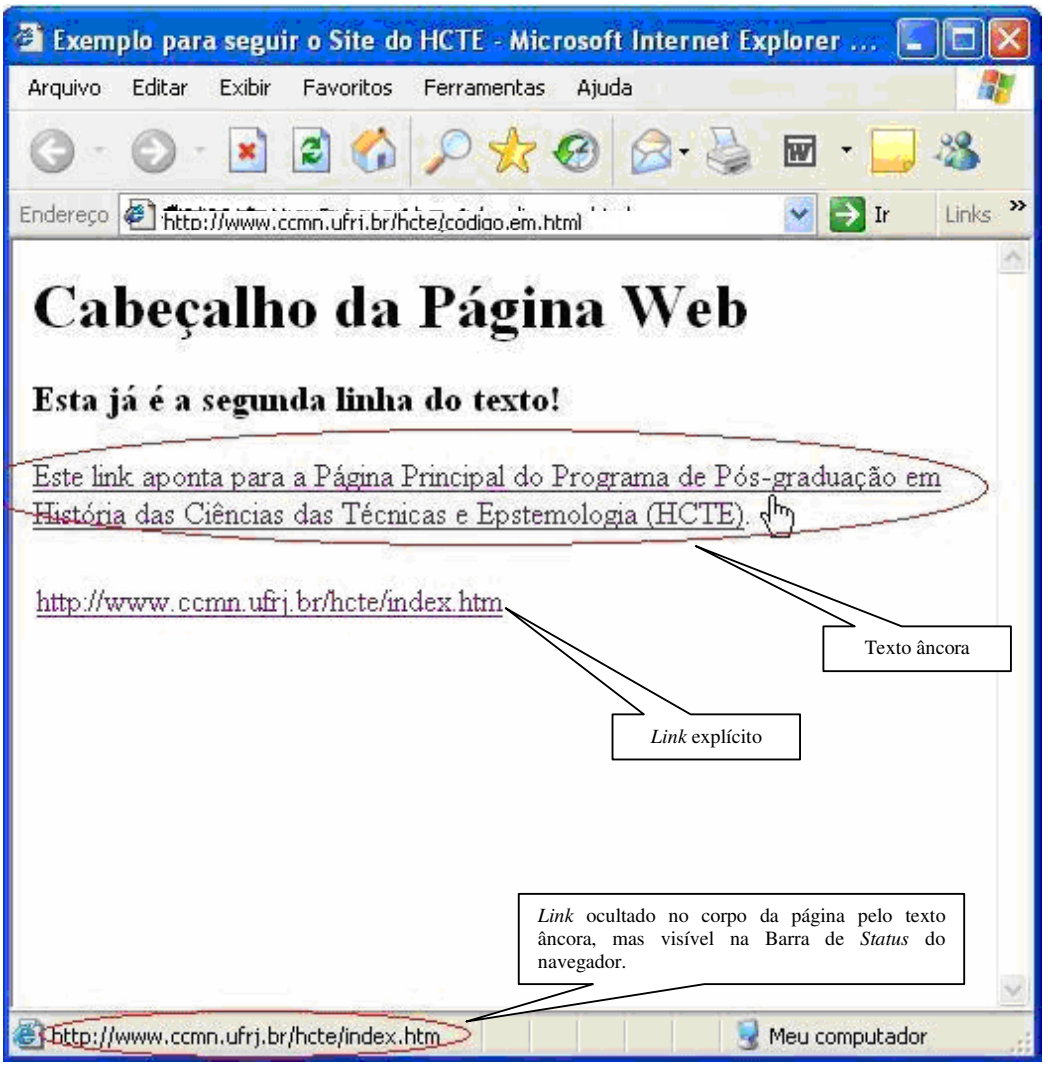

Figura 7 - *Links* e texto âncora

A frase "*Este link aponta para a Página Principal do Programa de Pós-graduação em História das Ciências, das Técnicas e Epistemologia (HCTE)*" é o texto âncora do referido *link* para a Página, ou neste caso o *site* do HCTE. Este texto é também utilizado pelos mecanismos de busca como um critério de relevância para uma página *Web* (ver capítulo 1). Se o texto âncora de um *link* contiver uma dada palavra, a página referenciada por esse *link* será tratada como relevante para esta palavra.

Segundo os fundadores da Google Inc., Brin e Page (1998), a sua ferramenta de busca na *Web* "emprega um grande número de técnicas para aumentar a qualidade da busca, incluindo, page rank (*sic*), texto âncora, e proximidade de informação". A verificação e utilização do "texto âncora" de um *link*, bem como as palavras do próprio texto da página *Web* sempre foram critérios utilizados abertamente pela Google Inc. na definição de sua ferramenta de busca. Associados ao *PageRank*, podem no entanto, levar a resultados controvertidos que passaram a ser chamados de *Google Bombs* ou *Google bombing*[49], que serão descritos seguir.

Normalmente, as causas de um *Google Bomb* são locais e/ou regionais, mas suas influências são percebidas em toda a Internet, pois a maioria dos mecanismos de busca têm abrangência mundial.

Um exemplo clássico e de amplo conhecimento, na área, de um *Google Bomb* foi o caso da expressão "*more evil than satan himself*". Ao ser colocada na ferramenta de busca como consulta, resultava como *hotspot*[50] a página principal da empresa Microsoft[51]. Este fenômeno atingiu não só o Google, mas também outras importantes ferramentas de busca na época (SPRING, 1999). Outro *Google Bomb*, com caráter político, que causou críticas diversas em abril de 2005, retornava como resultado para uma busca pela expressão "*miserable failure*" ou a palavra "*failure*" o *link* para a página

---

49 Esta expressão foi cunhada por Adam Mathes.

50 O primeiro *link* trazido na página da busca que é apresentada é considerado pela ferramenta como sendo o que tem maior relevância para aquela pesquisa feita pelo usuário e é denominado *hotspot*.

51 http://www.microsoft.com

Web da biografia do presidente norte-americano George W. Bush no *site* na Casa Branca, (como em destaque na Figura 8)[52]

Figura 8 - O *Google Bomb* – "*Failure*"

As causas centrais do surgimento do *Google Bomb* estão intrinsecamente ligadas aos critérios de funcionamento da ferramenta de busca adotados pela Google Inc. Podemos destacar três fatores como causas principais da formação de um *Google Bomb*:

- A utilização do critério de ranqueamento por popularidade e as características do algoritmo do *PageRank*, o qual considera o número de *links* apontando para uma página *Web* como fator de alta relevância para posicioná-la no seu *ranking*;

---

52 A imagem foi retirada de http://www.jabits.net/blog/PermaLink,guid,37348305-f8c2-41a6-8917-f583b7980954.aspx em 11/04/2006.

- A utilização deste mesmo algoritmo (*PageRank*) juntamente a técnica que se vale do texto âncora para o re-posicionamento da relevância da página, e que provavelmente dá pouca ou nenhuma importância a uma verificação explícita do conteúdo desta página para este ranqueamento;
- A intenção de indivíduos ou grupos de disseminar uma idéia visando mobilizar um grande número de pessoas em torno de um objetivo comum. Trata-se, como define Tatum (2005), de um exemplo da ação coletiva *online*.

2.2.2. Os *Blogs* e o *Google Bomb*

Para entender como se dá, na prática, a disseminação de uma idéia que possa vir a resultar em um *Google Bomb*, é preciso descrever o funcionamento de um recurso muito utilizado e difundido hoje para publicação de informação na *Web* – o ***Blog*** (*WebLog* ou *Web Log*)[53]. Esta ferramenta permite a um usuário da Internet, mesmo tendo poucos conhecimentos técnicos em codificação HTML e outros requisitos que seriam necessários para colocar uma informação *online*[54], o faça sem maiores dificuldades. É necessário apenas que este usuário se cadastre em um provedor – isto

---

53 Esta ferramenta de divulgação já é usada desde 1996, mas foi a partir de 2001 que se popularizou como diário pessoal *on-line* e público.

54 O uso do termo *online* (ou *on-line*), passou a ser utilizado na Internet, bem como na *Web*,como sinônimo de Virtual (Antunes e Correia, 2003), trazendo consigo a importância deste espaço público de difusão de informação. Lawrence (2001), do NEC Research Institute, relata o potencial de crescimento na comunicação entre os cientistas e pesquisadores e conseqüentemente avanço científico na utilização deste sistema facilitador.

pode ser feito sem custos pois muitos deles são gratuitos – que disponibiliza ferramentas para este fim e seguir alguns procedimentos simples[55]. Desta forma, poderá ser montado um texto formatado com letras de vários tipos, tamanho e cores, bem como incluir fotos para ilustrá-lo e criar *links* para outros *Blogs* ou *sites*. O *Blog* é usado com várias finalidades: um diário pessoal *online*, um fórum de discussão, divulgação de notícias e informações de todo tipo, onde há a possibilidade de interação entre o proprietário do *Blog* e seus leitores. Os *Blogs* normalmente têm um caráter regional ou local e abordam assuntos específicos de interesse do seu criador e/ou mantenedor. Podemos dizer que um *Blog* não é uma novidade como uma funcionalidade disponível na *Web*, pois não passa de um serviço que lança mão de recursos já bem conhecidos neste espaço, tais como páginas pessoais e os fóruns de discussão. Possivelmente, a novidade esteja na maneira como estes estejam sendo usados, permitindo disseminar informações e/ou opiniões individuais ou pessoais a respeito de qualquer assunto, desde política ou economia, até futebol ou pesquisas científicas. Outra característica é a velocidade em que os *Blogs* são atualizados e criados.

Haag (2006) ao comentar a importância dos *Blogs* na divulgação de informações, destaca a mudança do comportamento jornalístico, seja televisivo ou impresso, devido à velocidade da publicação de notícias por este meio. Alerta, também, para os riscos em relação à confiabilidade das informações e de suas fontes, por serem publicada por usuários muitas vezes leigos em jornalismo e que simplesmente expressam opiniões.

---

[55] Há muitas empresas que disponibilizam este recurso na *Web.* Os procedimentos de inclusão de usuários e das informações destes variam de empresa para empresa. Neste trabalho não haverá detalhamento do recurso por não ser este o objeto de estudo.

Devido às facilidades de publicação em um *Blog*, estes tornaram-se os principais disseminadores de um *Google Bomb* (SULLIVAN, 2002b). Se uma idéia ou um fato com capacidade de sensibilizar um amplo grupo ou numerosos indivíduos é lançado na rede, basta os usuários interessados, que controlam um *WebBlog* ou uma página *Web,* introduzirem um *link* com um texto âncora específico para a página alvo e estará disparado o gatilho de mais um *Google Bomb*. Como o *PageRank* dá prioridade ao número de *links* entre as páginas em detrimento do seu conteúdo (vide capítulo 1), a página referida passa a ter um alto valor (de *PageRank*), e quando associado àquele texto âncora, passa a ter relevância para o mesmo. Por fim, como resultado, a página visada pode passar a ser referenciada como *hotspot* na apresentação da busca para aquele texto âncora.

2.2.3. O posicionamento da Google Inc. frente ao *Google Bomb*

Embora tenham sido apresentadas algumas características do *Google Bomb*, não temos o objetivo de fazer um estudo aprofundado deste, mas sim destacar como a Google Inc., ao oferecer um serviço baseado em critérios de popularidade, possibilita o surgimento deste "fenômeno".

A Google Inc. tem-se mantido, até certo ponto, omissa no tocante a este problema. Mayer[56] (2005), em resposta ao aparecimento do evento do *Google Bomb*, argumenta que esta é apenas mais uma característica da ferramenta que sua empresa

---

[56] Marissa Mayer, Director of Consumer Web Products da Google Inc.

desenvolve, justificando que seria mais maléfico se a Google Inc. manipulasse diretamente os resultados das consultas para bloquear ou minimizar a ocorrência deste fenômeno. Por outro lado, a partir de 2005, a maioria dos resultados ligados a *Google Bombs* teve seus efeitos amenizados, provavelmente por mudanças relacionadas a alterações efetuadas pela Google Inc. nos algoritmos que compõem o *PageRank*. Pode-se então, concluir que, muitas vezes, os problemas que se apresentam ao usuário da *Web* são frutos não tanto de limitações técnicas, mas de escolhas feitas pela empresa em função de seus interesses os quais não são assumidos e colocados a público.

Byrne cita que:

> "O bombardeio ao Google (…) tem o potencial de levantar o debate e promover discussões através da Web, que pode desenvolver a dinâmica, temporal e descentralizada da natureza desta"[57]. (BYRNE, 2004).

Como este autor, acreditamos que o aparecimento do fenômeno do *Google Bomb* traz à tona discussões a respeito não apenas das características técnicas implementadas nas ferramentas de busca, mas também de aspectos ligados à própria constituição da *Web* e a ausência de normas pelas quais esta é regida, que nos parece pouco discutido pela sociedade e pelo seu principal interessado - o usuário.

---

[57] "Google Bombed (…) have the potential to raise debate and promote discussion across the web that can only further promote the dynamic, temporal and decentred nature of the web."

2.3. A privacidade do usuário: um "negócio da China"

Quanto ao problema das possibilidades do uso do mecanismo de busca da Google Inc. para ter acesso a informações privadas que foram inadvertidamente e/ou descuidadamente disponibilizadas na *Web,* tais como relatórios contábeis de empresas, currículos ou dados pessoais, bem como trabalhos que envolvam direitos autorais, cabe destacar que não se trata, necessariamente, de uma limitação ou falha da ferramenta, pois esta vasculha a *Web* atrás de todo tipo de informação. É necessário que o usuário esteja mais informado sobre os riscos inerentes ao sistema quando este busca usufruir dos recursos disponibilizados pelas empresas.

No caso dos registros que são gerados nos servidores *Web* da Google Inc., podemos afirmar que, mesmo esta não declarando que faz o cruzamento destes dados para fins de identificação do usuário, estas informações estão disponíveis para tal, como por exemplo o caso dos registros feitos quando um navegador solicita uma página *Web* ao servidor (veja o exemplo da Figura 9).

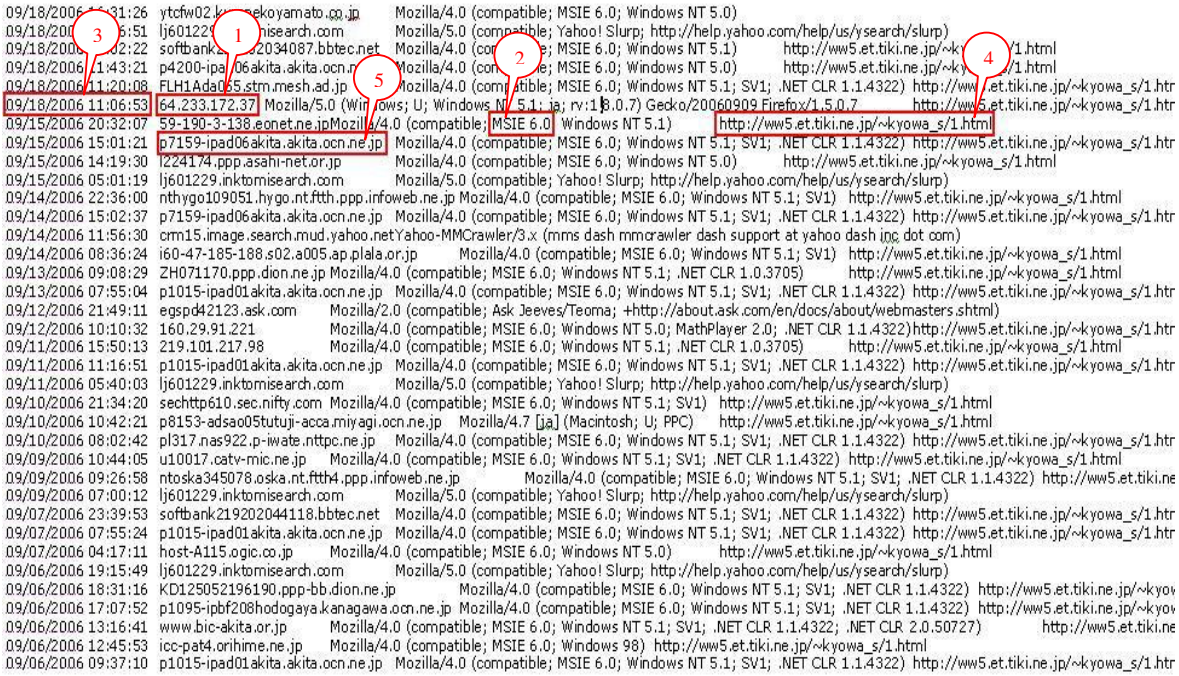

Figura 9 - Arquivo de registro de transação[58]

As informações mostradas na Figura 9 são geradas toda vez que um usuário faz um acesso a uma página de um servidor *Web*, através de seu navegador ou qualquer outro *software* que requisite um recurso disponível neste servidor, tais como uma

58 Esta imagem é parte de um documento disponível no *site* http://cgi.mediamix.ne.jp/~t5322/textlog100/log/log.txt, Estes registros foram conseguidos livremente através de uma simples consulta em uma ferramenta de busca (Google) do dia 23/09/2006. Até este momento não sabemos se estas informações foram disponibilizadas na *Web* intencionalmente ou por falha na segurança da informação desta empresa, pois os especialistas consideram estas informações de acesso restrito.

fotografia ou um filme. Serão destacadas apenas algumas das informações apresentadas[59], que bastarão para fundamentar nosso questionamento.

Como pode ser verificado no documento, está destacado o número IP[60] do computador do usuário (marcação 1 da Figura 9), que fez acesso através do navegador Internet Explorer 6.0 (marcação 2), a data e hora deste acesso (marcação 3) e, por fim, a página *Web* que foi acessada (marcação 4). Desta forma, se estas informações forem cruzadas com as consultas feitas pelo usuário na base de dados da ferramenta de busca, é então possível traçar um perfil detalhado de comportamento deste usuário, bem como a sua possível identidade. Esta característica não é exclusiva da ferramenta de busca da Google Inc., mas das ferramentas de busca de uma forma geral, bem como de empresas que disponibilizam páginas *Web* na Internet, ou que disponham de um servidor. Dificilmente o usuário poderá mascarar ou impedir que estes dados sejam capturados pelos sistemas da empresa.

Assim, um *site* especializado em música pode compartilhar dados de visitas e interesses de um usuário com uma empresa parceira que comercialize CD-ROMS de áudio. Desta forma, caso este mesmo usuário visite o *site* deste parceiro comercial, poderá receber propagandas de CD-ROM direcionadas aos seus interesses, que foram capturados no referido *site* de música. As informações podem, ainda, ser compartilhadas entre outras empresas parceiras comerciais. Voltaremos a aprofundar esta questão de

---

[59] Verificamos que são em torno de 20 tipos de informações registradas por um Servidor *Web*, que são comumente utilizadas pelos administradores do sistema.

[60] Como foi visto no capitulo anterior os números IP podem ser traduzidos para nomes (marcação 5 na figura 9) por servidores de DNS.

direcionamento de propaganda feita na *Web* por empresas de mecanismos de busca, no próximo capítulo.

Para exemplificar com um caso real, pode-se mencionar que houve nos EUA, em que a justiça norte-americana condenou à prisão perpétua um suspeito de assassinar a própria esposa. No processo foram utilizados alguns dados referentes a pesquisas realizadas pelo condenado empregando o mecanismo de busca da Google Inc., onde estavam registrados palavras como "pescoço", "pegar", "quebrar", bem como consultas à profundidade do lago em que o corpo foi jogado[61] (WALKER, 2006). Os dados foram cruzados com informações que o acusado tinha em seu computador, e que foram armazenadas pela ferramenta de busca da Google Inc. Este procedimento é normalmente utilizado entre as empresas que disponibilizam serviços na *Web*, através de pequenos arquivos chamados *Cookies*, tendo o objetivo de guardar informações sobre o usuário: desde uma conta de acesso, com a senha, até informações que permitam inclusive traçar a trajetória pelos *sites* que este mesmo usuário fez em um dado período, i.é, um histórico do que este usuário fez ou visitou em um respectivo *site*

Entendemos como Van Wel e Royakkers (2004), que as pessoas deveriam estar mais bem informadas sobre questões relativas à privacidade da sua navegação na *Web*, bem como dos seus possíveis usos para fins comerciais.

Em caso recente com o governo da China, a Google Inc. declarou que:

---

61 "*Forgot What You Searched For? Google Didn't*", artigo do The Washington Post: http://www.washingtonpost.com/wp-dyn/content/article/2006/01/20/AR2006012001799.html

> "mantém informações das pesquisas das buscas de seus usuários, em conjunto com os endereços IP (Internet Protocol) associados às consultas. O objetivo é melhor entender como seu mecanismo é usado..."(PETER NORVIG[62] apud MCMILLAN, 2006)

A empresa diz que armazenará estas informações fora do país asiático, para que os dados de seus usuários não possam ser violados, seja por este governo ou pelo governo dos EUA – país de origem da empresa. Uma outra demonstração de preocupação com a privacidade dos seus usuários ocorreu em janeiro de 2006, quando a Google Inc. ignorou, pela segunda vez, a intimação do governo norte-americano para enviar dados dos usuários de sua ferramenta de busca nos EUA, com o intuito de coibir atos de pedofilia na *Web* (Folha Online, 2006).

Por outro lado, segundo declarações da empresa para estabelecer-se na China, a Google Inc. passou a filtrar os resultados de consultas feitas por usuários deste país à sua ferramenta de busca, para que esta não apresentasse resultados contrários aos interesses no governo chinês. Poder-se-ia afirmar que esta mudança de postura foi influenciada por fins comerciais, pois "o país asiático tem cerca de 111 milhões de internautas - mercado que representa uma boa fonte de lucros para o Google"[63]

É possível perceber a dubiedade da empresa quando a questão é a defesa dos seus interesses e dos direitos dos seus usuários à privacidade e à liberdade de acesso à informação. Como mostram os exemplos acima, a Google Inc. delimita arbitrariamente,

---

62 Peter Norvig é diretor de pesquisa da Google Inc.

63 Folha Online – http://www1.folha.uol.com.br/folha/informatica/ ult124u19559.shtml. em 26/01/2006.

em função de seus próprios interesses comerciais, os direitos dos usuários ao acesso à informação. Ora o direito à privacidade e acesso à informação é garantido como no caso da contenda com a justiça norte-americana e posteriormente até com a justiça brasileira, ora os direitos a liberdade de acesso à informação é restringido, como no caso do governo chinês.

2.4. Considerações finais

Partindo das recomendações de Brant (2002), sobre as diretrizes que a ferramenta da Google Inc. poderia tomar para tornar-se, realmente, de "utilidade pública", citamos:

> "We feel that PageRank has run its course. Google doesn't have to abandon it entirely, but they should de-emphasize it. The first step is to stop reporting PageRank on the toolbar. This would mute the awareness of PageRank among optimizers and webmasters, and remove some of the bizarre effects that such awareness has engendered. The next step would be to replace all mention of PageRank in their own public relations documentation, in favor of general phrases about how link popularity is one factor among many in their ranking algorithms. And Google should adjust the balance between their various algorithms so that excellent on-page characteristics are not completely cancelled by low link popularity." […](BRANT, 2002)

A sugestão de retirada da apresentação do valor de *PageRank* na barra de ferramentas foi acatada pela Google Inc (ver Figura 10) nas versões atuais, apesar do motivo não ter sido divulgado[64]. Entretanto, acreditamos que as outras propostas citadas são de difícil implementação.

Figura 10 - Barra de Ferramentas da Google Inc. sem o valor do *PageRank*

Neste trabalho não pretendemos propor soluções rápidas e/ou simples a um problema tão complexo como é o caso da busca de informação na *Web.* No entanto, acreditamos que balizar os esforços apenas em um encaminhamento ligado à popularidade, como praticado pela Google Inc., pode levar a um fortalecimento exclusivo de grandes empresas em detrimento de grupos pequenos e novos (CHO e ROY. 2004), que não têm poderio econômico suficiente para terem uma boa classificação nos resultados das ferramentas de busca.

Assim é necessário um alerta para frear os exageros das promessas feitas pelas empresas que se propõem à realização dos desejos e necessidades do usuário, assim como o discurso de isenção sobre os resultados obtidos que as mesmas veiculam. É importante, neste momento, chamar a atenção para a necessidade de esclarecer os critérios técnicos e os interesses envolvidos na estruturação das ferramentas, visando

---

64 Será abordada outras questões ligadas às barras de ferramentas e seu uso comercial no Capítulo 3.

não induzir o usuário a erros. A *Web* é uma fonte de informação que cresce a uma velocidade impressionante e, cada vez mais, toma espaço na formação educacional e sócio-cultural da população.

Capítulo 3

AS EMPRESAS DE FERRAMENTA DE BUSCA – ASPECTOS COMERCIAIS

Neste capítulo, pretendemos esclarecer como algumas empresas e organizações de mecanismos de busca na *Web* se tornaram muito rentáveis. Buscaremos compreender como os maiores e mais conhecidos mecanismos de busca se mantém, crescem e trazem lucros para os seus acionistas, uma vez que as empresas responsáveis pelos mesmos divulgam que suas ferramentas e principais serviços podem ser usados sem necessidade do pagamento de qualquer taxa. Com este objetivo focaremos nosso estudo nas estruturas criadas pela Google Inc, empresa que detêm mais de 50% de todas as pesquisas realizadas em mecanismos de busca na *Web*[65].

Buscaremos, desta forma, analisar a lógica de sustentação financeira das empresas de ferramentas de busca e como elas estão articuladas à veiculação de propagandas na *Web.*

As demais organizações que oferecem serviços e/ou produtos no mercado local, regional ou mundial estão hoje, utilizando a *Web* como uma forma de divulgação e até mesmo, como meio de efetivação de transações comerciais e financeiras. Elas criam *sites* que apresentam seus produtos e serviços e indicam a forma de realização das transações.

---

[65] Fonte: Nielsen/NetRanting MegaView, janeiro 2007, em http://www.nielsen-netratings.com/pr/pr_070228.pdf.

É neste contexto que as ferramentas de busca passam a ter um papel fundamental na divulgação destas empresas, uma vez que o usuário do mecanismo de busca ao utilizá-lo para os mais diversos fins, passa a ser, também, um potencial consumidor das empresas que divulgam e/ou comercializam seus produtos e serviços na *Web*.

Verificamos uma profusão de novos serviços oferecidos na *Web*, especialmente, pela Google Inc. Segundo Jeanneney (2006) este "apetite gargantuesco" em apresentar novos produtos está vinculado à necessidade de atender à demanda de ganhos de seus acionistas uma vez que estes serviços atraem usuários que são os instrumentos de lucratividade da empresa.

3.1. Propaganda direcionada: funcionamento

As informações dos usuários são direta ou indiretamente coletadas pelas empresas de ferramentas de busca e são utilizadas como mercadoria na "sociedade da informação" ou "da economia da informação" (DEMO, 2000. Estas informações são capturadas automaticamente pelas ferramentas de busca em troca da oferta de serviços ditos "gratuitos". Este "comércio" é considerado por Battelle (2006) o carro-chefe dessas empresas de ferramentas de busca na *Web*. Cabe, aqui, a comparação com empresas de propaganda que sorteavam "*Ferraris"* para obter dados de potenciais consumidores, que ávidos pela oportunidade, ofereciam informações pessoais, muitas vezes sem saber que estas seriam vendidas a empresas interessadas em traçar perfis de consumo e explorar o mercado.

No caso das empresas de ferramentas de busca, a questão nos parece mais crítica, pois assim como vimos no capítulo anterior, elas retêm informações privilegiadas e

mais completas de seus usuários, tais como locais onde vivem, trabalham, horários, lazer, desejos e interesses, os quais por ventura estejam pesquisando na *Web*. Estes dados são coletados sem que os usuários estejam devidamente informados sobre o procedimento e a utilização de suas informações. A empresa mantém em um link interno em sua página "um contrato de uso do serviço" que afirma, apenas, que estes dados não serão vendidos diretamente a parceiros ou a outras empresas e, que o usuário que não concordar com o procedimento poderá não ter acesso ao serviço "gratuito" oferecido pela empresa. A observação na página do Google menciona que:

> "Você pode se recusar a submeter as informações pessoais a qualquer um de nossos serviços, caso em que haverá a possibilidade do Google não ser capaz de lhe fornecer esses serviços.
> You can decline to submit personal information to any of our services, in which case Google may not be able to provide those services to you." [66] (GOOGLE, 2005)

A justificativa dada pela empresa para a apropriação das informações dos usuários é de que estas serão utilizadas para melhoria dos serviços prestados. A empresa menciona que:

> "Podemos combinar a informação que você submete em sua conta com a informação de outros serviços do Google ou

[66] O Google utiliza, freqüentemente, um tradutor automático para as suas páginas oficiais que não estejam em Inglês, entretanto, como ressalta Jeanneney (2006), este tradutor funciona precariamente. No caso desta citação, a própria página do *site* da empresa traz a versão original (em inglês) e logo abaixo a tradução.

> terceiros a fim de fornecer-lhe uma experiência de navegação melhor e para melhorar a qualidade de nossos serviços." (GOOGLE, 2005)

As empresas de ferramentas de busca após capturarem as informações de seus usuários procuram associá-las a produtos ou serviços de outras empresas que estejam dispostas a pagar por esta intermediação. Para melhor compreender esta inter-relação é necessário compreender o conceito e funcionamento de "palavra-chave" no comércio entre as empresas na *Web*.

3.1.1. Palavras-chave: *Link* x *Banners*

Como foi visto no capítulo 2, o termo palavra-chave é relacionado ao funcionamento dos mecanismos de busca em suas etapas de coleta, ordenação e apresentação de resultados. Ao fazer uma consulta, o usuário insere uma palavra ou frase em uma ferramenta de busca. Ele espera obter algum resultado que o permita encontra uma página na *Web* com o assunto desejado. No entanto, para a lógica comercial da empresa de ferramenta de busca, a utilização de uma dada palavra pelo usuário o identifica também, como um potencial consumidor de produtos e serviços, que a ela podem ser associados. Desta forma, estas empresas abrem um espaço para o comércio com as palavras-chave.

Assim, no momento em que o usuário faz a pesquisa, dois mecanismos são acionados: um que traz os *links* das páginas *Web* associadas aos termos digitados e outro que apresenta *links* que são propagandas de clientes ou parceiros das empresas de

ferramentas de busca. Ambos os resultados são apresentados ao usuário em uma mesma interface da ferramenta de busca. As empresas desta área distinguem os primeiros como resultados orgânicos do sistema automático e os outros como "*Links* Patrocinados" (BATTELLE, 2006) (veja figura 12). Este mecanismo abre imensos espaços e já movimenta bilhões de dólares anuais.

A propaganda direcionada veiculada nas ferramentas de busca na *Web* pode ser apresentada ao usuário em uma página *Web* de duas formas, através de *Banners* ou *Links* Patrocinados.

Um *Banner* ou *Web Banner* é a forma de propaganda apresentada, normalmente, na parte superior e mais visível da página, contendo cores e figuras, muitas vezes em movimento, objetivando chamar a atenção do usuário. O conteúdo pode estar vinculado ou não à pesquisa feita pelo usuário, porém é de fácil reconhecimento que se trata de uma propaganda. Por outro lado, a propaganda direcionada ou *links* patrocinados, dissimulada em *links*, apresenta pouca poluição visual, que em alguns casos, pode confundir o usuário, levando-o a não distingui-las dos resultados orgânicos apresentados pela ferramenta de busca.

As palavras-chave, ao permitirem a associação de produtos e serviços às buscas que os usuários realizam na *Web*, passam a adquirir uma grande importância comercial e a serem oferecidas pelas empresas de mecanismo de busca através de leilões para serem disputadas. As empresas que fazem os melhores lances podem ter o seu serviço ou produto associado a elas quando estas forem digitadas pelo usuário da ferramenta.

Os leilões na maioria das vezes são realizados através de sistemas automatizados, aumentando com isso a agilidade do processo e possibilitando a

maximização do valor alcançado pelas propostas de uso de cada palavra, uma vez que um número maior de anunciantes pode participar da disputa. Os leilões estabelecem o preço a ser pago pelo "*click*". A explicação deste último termo será dada na seção a seguir.

3.1.2. *Pay-per-click*

A utilização de palavras-chave em uma campanha publicitária está, geralmente, associada a um mecanismo chamado de *pay-per-click* (PPC). O anunciante que detém o seu produto ou serviço articulado à palavra, paga apenas se o usuário exposto à propaganda clicar no *link* especificado. Desta forma, há o pagamento apenas se houver algum interesse deste usuário - potencial consumidor - pelo produto ou serviço oferecido. Ou seja, será cobrado do anunciante apenas se o usuário clicar sobre o *link* patrocinado. Este idéia foi primeiramente lançada por Bill Gross através de sua empresa Idealab! e posteriormente comercializada através da Overture (BATTELLE, 2006).

Atualmente, há muitas outras variações desta mesma prática sendo utilizadas por diversas empresas de mecanismos de busca. A Google Inc. nomeou o seu sistema de PPC de *AdWord* e o incorporou à vários de seus serviços, como às suas ferramentas de busca e ao Gmail[67]. Este sistema, além de fazer a associação entre a palavra digitada pelo usuário de um dos serviços com *links* patrocinados, permite controlar se estes *links* apresentados foram ou não acessados. É importante ressaltar que, por se tratar de um

[67] http://ww.gmail.com, o serviço "gratuito" de correio eletrônico da Google Inc.

sistema de uma empresa em grande expansão, novas funcionalidades vão, constantemente, sendo agregadas.

Na figura 11, é apresentada parte de uma página de resultados de uma pesquisa com a palavra "book", utilizando o mecanismo de busca Google. Nesta figura, destacamos o que obtemos acima da posição de *hotspot*: o item que é chamado de "*Link* Patrocinado". Como visto, trata-se de uma propaganda associada à palavra "book", resultado do sistema *AdWord* da empresa. É importante ressaltar que o *link* ou *links* patrocinados, freqüentemente, ocupam o lugar do *hotspot* na página de resultados da pesquisa.

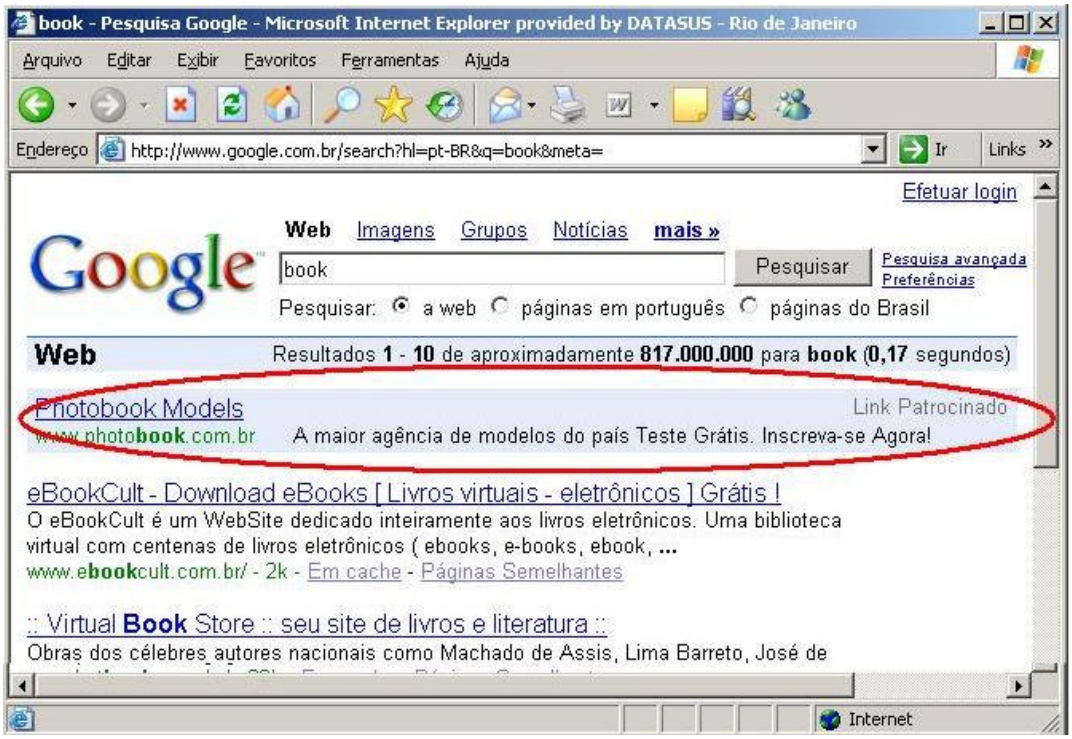

Figura 11 – Parte da página de resultados de uma pesquisa com a palavra "book" no Google.

A figura 12 apresenta uma página completa de resultados de uma pesquisa com a palavra "livro", utilizando o mecanismo de busca Google. Pode-se notar que há outras posições na qual a propaganda direcionada ou *links* patrocinados podem ser apresentados na página de resultados de uma pesquisa. Foram feitos destaques em vermelho, nesta figura 12, para melhor identificar os espaços de propaganda mais utilizados.

Figura 12 – Página de resultados de uma pesquisa com a palavra "livro", utilizando o Google.

Nestes dois exemplos é importante verificar, também, que o navegador está direcionado para a página do Google no Brasil, na qual as propagandas estão redirecionadas regionalmente. É possível notar que, dependendo da palavra-chave ou expressão digitada, podem existir mais de uma propaganda, mostrando, desta forma, que há termos mais disputados, comercialmente, na *Web*, do que outros.

Há outra variação de propaganda direcionada que também lança mão do mesmo princípio do PPC. Trata-se do mecanismo que redireciona a propaganda de um cliente X (anunciante) para a página de outro cliente Y. Normalmente isto é feito se Y tem uma página que é muito visitada, Assim X, ao invés de fazer a propaganda na página do próprio Google, passa a fazê-la na página de Y e a Google Inc. funciona somente como intermediária, isto é, intermedia o negócio entre X e Y. A Google Inc. desenvolveu o *AdSense*, através do qual o interessado em comercializar um espaço para propaganda em sua própria página, pode se inscrever neste sistema. O proprietário da página recebe à medida que seus "visitantes" clicam no *link* da propaganda. Verifica-se que para garantir um maior interesse, os anúncios são associados ao conteúdo das páginas *Web* que os apresentam.

3.1.3. SPAMMERS e SEOs

O aumento de popularidade e de exposição que uma empresa ao ser listada nos resultados de um mecanismo de busca deu origem a um novo comércio de serviços na *Web* - os SEOs. Como vimos anteriormente, os SEOs são empresas especializadas em fazer com que páginas ou *sites* sejam incluídos ou alcancem melhor posição nos resultados das pesquisas oferecidos pelos mecanismos de busca. Com esse objetivo,

lançam mão de muitas técnicas que as próprias empresas de ferramenta de busca tentam neutralizar.

As técnicas utilizadas pelas empresas de SEOs vêm sendo desenvolvidas a partir de análises do comportamento das ferramentas de busca e são baseados em alguns critérios explicitados por elas, alguns dos quais foram descritos no capítulo 1 deste trabalho.

Quando fica caracterizado que uma dessas empresas conseguiu alterar significativamente os resultados de uma pesquisa em uma ferramenta de busca, a empresa que controla a ferramenta pode punir a empresa responsável excluindo-a de sua base de dados. Como constatamos em nossa pesquisa, a Google Inc. impõe ao considerado infrator a penalidade de redução do *PageRank* de sua página em seu sistema, imputando-o o valor igual a zero, causando a assim, a exclusão da página do ranqueamento, e consequentemente do processo de coleta de seus robôs.

A prática de adulteração de resultados orgânicos de uma ferramenta de busca é considerada um *Spam* (*Web search engine spam*) e seu praticante um *Spammer*. Esta mesma expressão é conhecida e usada também na prática de envio de mensagens de correio eletrônicos em massa para destinatários que não solicitaram ou autorizaram este envio.

Barras de Ferramentas, como visto no capítulo anterior, adicionam funcionalidades aos navegadores *Web*, tais como o Internet Explorer, Firefox e Netscape. São também um recurso utilizado por muitas empresas de ferramenta de busca (Google, Yahoo!, Mamma) para aumentar a popularidade de seus serviços e o

acesso a seus *sites.* Por exemplo, através da barra de ferramentas, o usuário pode ter acesso direto ao mecanismo de busca do Google, reforçando o uso deste último.

Temos também a valorização do *hotspot* e do posicionamento dos *links* na primeira página de resultados, inclusive com a utilização de recursos, como por exemplo, o "Estou com Sorte" ("*I´m feeling luck*") da Google Inc.. Trata-se de um botão na página principal do Google que, ao ser acionado, apresenta apenas o *link hotspot*, eliminado ou ocultando os outros *links* no resultado pesquisa. Isto pode denotar uma intencional supervalorização do espaço disponibilizado pelas ferramentas de busca, uma vez que 80% dos usuários examinam apenas os dez primeiros resultados da pesquisa apresentada por uma ferramenta de busca (EASTMAN e JANSEN, 2003).

Na "guerra" travada entre proprietários de *sites* para tornar-se ou manter-se no topo da primeira página de resultados de uma pesquisa, diversas estratégias são utilizadas, desde o monitoramento dos critérios de busca e indexação de uma empresa até a manipulação das páginas rivais.

Neste universo, os interesses comerciais que influenciam a ordenação dos resultados apresentados nas pesquisas podem ser considerados legítimos, desde que o usuário esteja devidamente informado sobre o procedimento.

Cabe mencionar que três bilhões de consultas mês foram executadas no Google, em dezembro de 2006, equivalente a 51% de todas as consultas computadas pela NetRanting Inc.[68] neste período. Tendo em vista que para cada consulta efetuada é

---

[68] http://www.nielsennetratings.com.

apresentada uma página de resultados similar a dos exemplos das figuras 11 e 12, conclui-se que a quantidade de propaganda paga veiculada por esta ferramenta tem uma expressão significativa. Estar em primeiro lugar, ou mesmo, na primeira página, significa ter visibilidade entre bilhões de potenciais consumidores.

O grau de influência desses resultados sobre o usuário pode ser balizado pela pesquisa conduzida por Fallows (2005). Ela mostra que 68% dos usuários entrevistados dizem que os mecanismos de busca são fontes seguras e não tendenciosas de informação, sendo que apenas 19% não acreditam nisso. Desta forma, verifica-se a importância de se debater as questões do patrocínio dos serviços de ferramentas de busca e outros mecanismos ainda mais obscuros que interferem nos resultados apresentados ao usuário. Conforme declarado pela Google Inc.:

> "*A Google search is an easy, honest and objective way to find high-quality websites with information relevant to your search.*" (GOOGLE)

A utilização da popularidade como principal critério de relevância para indexação de páginas na *Web*, pode causar favorecimento dos grandes *sites* em detrimento dos menores e recém-criados, independente do conteúdo e qualidade destes (CHO e ROY. 2004 e INTRONA e NISSENBAUM, 1998). No entanto, Fortunato *et al* (2005), da Universidade de Indiana (EUA) em uma pesquisa baseada em dados estatísticos e na utilização de ferramentas (*softwares*, base dados e infra-estrutura) de alguns dos mais conhecidos mecanismos de busca norte-americanos - Google, Yahoo!, Altavista (atualmente pertencente à Yahoo!) e Alexa (Amazon) - comprovam que, se não existissem mecanismos de busca automáticos na *Web*, os *sites* menores e mais

novos teriam ainda mais dificuldades de se tornarem populares. Segundo os autores, os pequenos se beneficiam em maior grau dos mecanismos de busca do que os *sites* grandes e populares. De acordo com a pesquisa, os mecanismos de busca favorecem o trafego para estes novos *sites* ultrapassando o número de acessos que os mesmos receberiam se estes fossem feitos apenas via navegação por *links* na *Web*.

Entretanto, assim como Lawrence e Giles (1999), entendemos que a existência e atuação dos mecanismos de busca na *Web*, privilegiando o critério da popularidade tendem cada vez mais a favorecer as páginas mais populares e muito acessadas, enquanto que as novas páginas têm dificuldade ainda maior de aparecerem nos resultados destes, independente da qualidade de seus conteúdos.

Apesar de parecer um pouco exagerada, uma das "máximas" deste mercado é "se a sua página não estiver indexada pelo Google, ela não existe na *Web*" (CHO, ROY e ADAMS *apud* OLSEN, 2005). Dentro desta problemática, torna-se necessário "saber quem vem primeiro – se o Google esta refletindo a popularidade ou está criando popularidade" [69] e quais os critérios e métodos utilizados pela empresa no seu processo de indexação de páginas.

Se os Spammers e SEOS são formas de interferir no ranqueamento, como podemos saber se as empresas de ferramenta de busca também não laçam mão de políticas de favorecimentos de empresas que interferem nos resultados oferecidos ao usuário? Elas afirmam que não podem divulgar seus critérios e métodos para que seus

---

[69] Daniel Brandt à Cnet.com, em http://news.com.com/2009-1023-963618.html

sistemas não sejam manipulados pelos Spammers e SEOS. No entanto esta ocultação, no mínimo, pode colocar em dúvida a legitimidade dos resultados obtidos quando buscamos informações nestas ferramentas.

3.2. Comércio de Inclusão na base de dados dos mecanismos de busca na *Web* (*pay-per-inclusion*)

Atualmente, muitas empresas de ferramenta de busca comercializam a inclusão de páginas em sua base de dados, como por exemplo, a Yahoo!. Inicialmente, esta ferramenta de busca tinha o formato de Diretório, como visto no capítulo 1, sendo que um de seus principais argumentos para a utilização desta prática era que, com o pagamento, o responsável pela página interessado na sua inclusão na base de dados desta ferramenta poderia ter suas informações mais rapidamente avaliadas pelos seus editores e consequentemente, mais rapidamente teria sua página apresentada no resultado das pesquisas. É importante ressaltar que, segundo a Yahoo!, apenas é vendida nesta modalidade a inclusão na base de dados e não o posicionamento desta página no resultado da busca feita pelos usuários. Desta forma, a indexação ou ordenamento fica associado a outros critérios de relevância que são empregados nas ferramentas de busca.

Algumas empresas de ferramenta de busca, principalmente, as automáticas, declaram que não utilizam esta prática de cobrança para inclusão. Dentre elas encontra-se a Google Inc. Desta forma, apresentação das páginas nestas ferramentas seria em função de critérios de inclusão e ordenação, e não o de pagamento. Procuraremos mostrar através de exemplos e descrições de práticas comerciais dessas empresas que,

apesar das declarações, estas empresas acabam por incluir os seus parceiros e clientes de seus serviços pagos na base de dados de suas ferramentas.

A Google Inc., como vimos anteriormente neste capítulo, vende espaço de propaganda em outros *sites*, intermediando a relação comercial entre anunciante e local de propaganda, isto ocorrendo através do seu sistema *AdSense*. Segundo Vise e Malseed (2005) empresas com sites grandes e populares na *Web*, como AOL, The New York Times e Univision, utilizam-se deste sistema da Google Inc. E, conforme discutido por Battelle (2006), a disputa entre a Overture e a Google Inc. pela parceria com a AOL, tendo como vitoriosa a segunda, mostra a acirrada disputa por parceiras com empresas deste porte na Internet.

Empresas que podem pagar para serem apresentadas em propaganda direcionada através de *links* em sites grandes e populares passam a ter mais chances de, não somente serem incluídas, mas também melhorarem seu posicionamento. O *PageRank*, como direcionador dos robôs do Google (*Googlebots*), não só prioriza as páginas populares mas também os *links* para as páginas que nelas estão, já que sua pontuação, feita pelo dito algoritmo, passa a ser mais alta, como vimos no capítulo 2.

Portanto, uma empresa que não possa comprar espaço de propaganda através do *AdSense* terá menos possibilidade de ter sua página ou *site* incluídos na base deste mecanismo de busca do que outra que, mesmo sendo nova e pouco popular, possa comprar estes serviços oferecidos pela própria Google Inc.

Como pode ser verificado, este é certamente um método para inclusão e também para melhora do posicionamento de uma página de resultados do Google.

Considerando a prática comercial desta empresa ao vender espaços de propaganda em *sites* grandes e populares e, associada à popularidade, o principal critério que norteia o algoritmo de sua ferramenta de busca - o *PageRank*, bastariam para concluirmos esta alegação.

3.3. Comércio do posicionamento dos *sites* na apresentação da pesquisa dos mecanismos de busca na *Web* (*pay-per-placement*)

A cobrança para o reposicionamento de uma página nos resultados de pesquisa de uma ferramenta de busca tornou-se uma prática, pela primeira vez, em uma ferramenta de busca chamada GOTO.COM em 1998, na qual da mesma forma que um serviço prestado pelas "Listas Amarelas", os interessados pagavam por sua inclusão, de tal forma que, quem pagasse mais teria sua página melhor posicionada no resultado da busca feita por esta ferramenta (PELLINE, 1998). A página com os resultados de uma pesquisa era complementada com pesquisas obtidas de outra ferramenta de busca - o Inktomi, que atualmente faz parte da Yahoo! Inc. (BATTELLE, 2006). A justificativa dos responsáveis pela ferramenta era de que um critério de relevância que considerasse também a capacidade da empresa de pagar por um espaço privilegiado, poderia servir como referência para a qualidade do produto ou serviço oferecido pela empresa em comparação com as demais.

Como exemplo, poderíamos comparar a prática do *pay-per-placement* (PPP) ao comércio de propaganda ou classificados de jornal convencional, onde além de pagar para ser incluído o anunciante pode escolher o posicionamento e tamanho do anúncio a ser publicado. Da mesma forma, uma empresa que deseja divulgar seu produto ou

serviço na *Web*, pode fazê-lo através de uma página de apresentação de resultados de uma ferramenta de busca. Caso o usuário esteja buscando algum produto ou serviço na Internet e disponha de tempo para comparação de qualidade e preços, os resultados oferecidos podem lhe favorecer. Entretanto, este usuário, muitas vezes, tem outras necessidades ou objetivos (SULLIVAN, 2001).

Esta prática não é considerada irregular ou ilegal, desde que os *links* patrocinados apresentados pelas ferramentas de busca venham discriminados como tal. No entanto, em 2001, houve uma reclamação à Federal Trade Commission[70] (FTC), sobre a inclusão de propaganda sem aviso ou descriminação entre os resultados das buscas de alguns dos mais populares mecanismos de busca na época: Microsoft, Altavista e AOL. Estas empresas receberam recomendações para que alterassem suas práticas com relação à exibição destes *links* patrocinados.

Nicholson *et al* (2006), apontam que 40%, em média, dos resultados apresentados pelos mecanismos de busca são propagandas pagas. Analisando a figura 12, no inicio deste capítulo, podemos verificar que mesmo o Google, que se valoriza pela discrição ao apresentar "*links* patrocinados", atinge, facilmente, este percentual. Cabe destacar que há variações neste índice, dependo do interesse comercial que cada palavra-chave apresenta.

Conforme destacam Friedland e Gelsier (2005), a partir de 2006, anúncios através de *links* patrocinados começam a substituir as outras formas de propaganda na

---

[70] Órgão responsável por promover a proteção do consumidor e práticas antitruste dos EUA.

Web. Afirmam também que, anunciantes gastarão 6,1 bilhões de dólares com *links* patrocinados e cerca de 6,4 bilhões em todas as demais formas de propaganda na *Web*: como *banners*, listas de classificados. Neste contexto, a IDG Now![71] prevê que "A partir de 2010, companhias devem gastar 17,3 bilhões de dólares em busca patrocinada, contra 12,3 bilhões de dólares em outros tipos de propaganda *online*. De 2005 a 2010, o crescimento anual estimado para *links* patrocinados e outras categorias é de 23% e 14%, respectivamente". Desta forma, tem-se com os *links* patrocinados um importante meio de propaganda na *Web*.

3.4. A utilização dos dados dos usuários e a privacidade

A Google Inc. lança mão da propaganda direcionada em quase todos os seus serviços "gratuitos". Além dos mecanismos já citados, quando do acesso do usuário a ferramenta de busca, um dos mais polêmicos empregos desta prática está no serviço de correio eletrônico oferecido pela empresa: o Gmail[72]. Quando o usuário acessa o seu correio eletrônico através de um navegador para ler suas mensagens, o sistema *AdWord* é acionado e vai em busca de palavras "comerciais" no texto da mensagem do usuário. Estas palavras-chaves são associadas a produtos e serviços de empresas que são parceiras ou clientes comerciais da Google Inc. Desta forma, o sistema apresenta propagandas na página de correio do usuário. Por exemplo, se o conteúdo da mensagem

---

71 http://idgnow.uol.com.br/internet/2005/08/08/idgnoticia.2006-03-12.8898521567

72 http://www.gmail.com

for sobre uma viagem, o sistema pode apresentar propagandas de parceiros ou clientes relacionados a este tema, tais como propostas de cruzeiros e estada em hotéis, que aparecem como *links* ao lado do texto das mensagens.

A utilização de propaganda é uma prática comum em serviços de correio eletrônico via *Web* (*webmail*), sendo utilizadas também pela Yahoo! (Yahoo!Mail), Microsoft (Hotmail), etc. Entretanto, a avaliação direta do conteúdo das mensagens é novidade. Apesar da alegação da Google Inc. de que o processo de "ler" as mensagens de seus usuários seja automático e através de um sistema que não o identifica, mas apenas faz uma "varredura automática"[73], entendemos que esta é um assunto muito polêmico e envolve questões éticas. Estas informações ficarão nos registros (*logs*) do servidor e aplicativos da empresa e balizam o seu comercio de propaganda, sendo que esta é uma das utilidades dada a elas que temos conhecimento hoje. Porém, não sabemos todas as utilidades a que elas servem ou servirão. Trinta organizações de direitos a privacidade de várias partes do mundo, em carta aberta[74] endereçada aos fundadores da Google Inc., argumentam inclusive, que não só são violados os direitos dos usuários cadastrados neste serviço, mas também quem envia mensagens para estes.

Esta questão da utilização da varredura automática como álibi para justificar a invasão de privacidade remete a uma série de questões muito interessantes que no entanto não teremos condições de aprofundar. Em primeiro lugar temos em jogo a questão da neutralidade da ciência e da tecnologia. Ou seja, a uma máquina é permitido

---

73 http://mail.google.com/mail/help/intl/pt-BR/about_privacy.html

74 http://www.privacyrights.org/ar/GmailAGadvisory.htm

vasculhar na intimidade de uma correspondência sendo o mesmo proibido para o ser humano. Como se houvesse diferença entre alguém ler um texto à procura da palavra corrupção, por exemplo, ou fazer um programa que assinale com um asterisco se o texto contiver a mencionada palavra. Qual a diferença? No fundo, o que está em jogo é fundamentalmente uma questão de concentração de poder. Se na primeira opção o poder fica descentralizado entre aqueles que lêem os textos na segunda opção há uma concentração naquele que lerá o relatório do computador.

O objetivo não é levantar questões jurídicas, mas questionar práticas como esta, que atingem diretamente a vida privada dos usuários, bem como indagar até que ponto os mesmos têm clareza das formas de utilização de suas informações e das possíveis conseqüências, uma vez que estas estão sendo, não apenas coletadas, mas centralizadas nas mãos de um pequeno grupo de empresas.

Finalmente, destacamos que as parcerias entre as empresas de mecanismos de buscas são pouco duradouras e desfazem-se tão facilmente quanto são estabelecidas. Como exemplo, pode-se citar o caso Google-Yahoo! entre 1998 e 2001, ou mesmo o atual Google-DMoz, sendo que as empresas da primeira parceria posteriormente se tornaram "rivais" declaradas, disputando o mesmo mercado de usuários e possíveis consumidores de mecanismos de busca. Inicialmente, a Google Inc. lançou mão dos resultados do Diretório do Yahoo! para alimentar sua lista de coleta dos seus robôs (*Googlebots*). Atualmente, ela tem seu próprio serviço de Diretório, que dispõe de resultados obtidos de outra organização - o DMoz - que também complementa os resultados do seu Mecanismo de Busca Automático.

Observamos que a dinâmica de associação e concorrência dessas empresas não se diferencia das demais do mercado cujos interesses econômicos prevalecem nas

tomadas de decisão. Podemos concluir que o objetivo destas negociações é melhorar o posicionamento da própria empresa no mercado de buscas na *Web* e não, necessariamente, ampliar a qualidade dos serviços prestados aos usuários de tais ferramentas, como reiteradas vezes elas divulgam. Como apresenta Demo (2000),

> "A sociedade da informação informa bem menos do que se imagina, assim como a globalização engloba as pessoas e povos bem menos do que se pretende. Na sociedade da mercadoria, mercadoria vem antes." (DEMO, 2000).

Capítulo 4

CONSIDERAÇÕES FINAIS

Conforme discutido nos capítulos anteriores, os mecanismos que as empresas de ferramentas de busca utilizam para organizar as páginas e estabelecer o ranqueamento que é oferecido aos usuários em suas pesquisas são limitados por fatores de ordem técnica, econômica e política.

Do ponto de vista técnico, a limitação advém do grande volume de informações existentes na internet, gerando a incapacidade por parte das empresas de coletá-las, ordená-las e oferecê-las ao usuário de forma acessível e democrática. A necessidade de atingir todo o espectro de informações disponível na *Web* é questionada por Diaz-Isenrath (2005) partindo do pressuposto que quantidade não necessariamente reflete qualidade. A autora questiona, ainda, se as informações "pertinentes", são, realmente, as mais "populares", como querem induzir as empresas que controlam o desenvolvimento das ferramentas de busca. É importante ressaltar que as escolhas técnicas adotadas normalmente têm um direcionamento dado por interesses econômicos dessas empresas.

Do ponto de vista político, temos a influência de governos que, desejando garantir o controle sobre a população de seu país, interferem, pressionando as empresas de mecanismos busca a adaptarem suas ferramentas de forma que seus interesses políticos sejam preservados. Essas empresas, em função de seus interesses de ordem econômica oscilam, ora garantindo o direito à privacidade e o acesso à informação - como no caso da contenda entre a Google Inc e a Justiça norte-americana, ora cedendo

às pressões, cerceando o direito de acesso à informação de seu usuário - como ocorreu com a empresa quando pressionada pelo Governo chinês.

Do ponto de vista econômico, cabe questionar a comercialização da inclusão e do ranqueamento de páginas na sua listagem de resultados das ferramentas de busca. Isto influencia, sobremaneira, o resultado das pesquisas obtidas pelo usuário. Por outro lado, essas empresas vêm comercializando, direta ou indiretamente, informações dos usuários, coletadas através de seus sistemas.

Cabe destacar que na "Sociedade da Informação" o conceito tradicional de mercadoria passa a ser expandido sendo que estas informações coletadas pelas empresas devem ser consideradas uma mercadoria e a Internet, um espaço para sua comercialização conforme Demo (2000). Desta forma, torna-se premente questionar o uso que as empresas de mecanismos de busca vêm fazendo das informações obtidas de seus usuários. Elas afirmam que os serviços oferecidos são gratuitos e em troca coletam e tomam de seu usuário dados que na verdade são produto de elevado valor no mercado. No caso da Google Inc., por exemplo, além de um contrato obscuro presente numa página interna da empresa, ela tenta justificar o seu desrespeito ao direito de privacidade do usuário afirmando que a apropriação destas informações é necessária para ampliar e melhorar os serviços oferecidos ao usuário.

Não é objetivo deste trabalho propor soluções definitivas ou emergenciais para o complexo problema que envolve os sistemas de busca de informação na *Web*. No entanto, acreditamos que sobre valorizar a "popularidade" – como é praticado pela Google Inc. entre outras, pode levar a um fortalecimento das grandes empresas em detrimento de grupos pequenos e novos, que não têm poderio econômico suficiente para terem uma boa classificação nos resultados das ferramentas de busca, conforme

ressaltado por Cho e Roy (2004). Também entendemos, como Van Wel e Royakkers (2004), que o usuário deveria ser melhor informado sobre as limitações do sistema, como por exemplo: a impossibilidade técnica de oferecimento democrático das informações disponíveis, o cerceamento da liberdade de acesso à informação, os riscos de quebra de privacidade presente na *Web*, bem como, dos usos comerciais que vem sendo feitos aos dados pessoais dos usuários. Assim, acreditamos que o questionamento das práticas e dos discursos das empresas de mecanismos de busca na *Web* deve partir do pressuposto que o objetivo primeiro destas empresas deve ser o de organizar a informação e permitir o acesso – o mais democrático possível - do conteúdo amplo e diversificado presente na Internet. Desta forma, o funcionamento das ferramentas de busca deve ser transparente, explicitando as estruturas técnicas, econômicas e políticas que influenciam no seu funcionamento.

BIBLIOGRAFIA

Amorim, Ricardo e Vicária, Luciana, 2006, "A Enciclopédia Pop", **Revista Época**, n.401, pp. 40-47, São Paulo.

Antunes, Mafalda e Correia, Susana. 2003. **Semantic nets in the Net**. Em: Proceedings of CIL17. Hajicová, Kotesovcová & Mírovský (orgs.). Praga, República Checa. Editora: Maftyzpress, MFF, UK. Também encontrado em http://www.iltec.pt/pt/handler.php?action=artigos&book=95, última visita 04/10/2006.

Arasu, Arvind, Cho, Junghoo, Garcia-Molina, Hector, Paepcke, Andreas, and Raghavan, Sriram. 2001. "Searching the Web". **ACM Trans. Inter. Tech. 1**, 1 August., pp. 2-43. DOI= http://doi.acm.org/10.1145/383034.383035.

Arnold, Stephen E. 2004. "How Google Has Changed Enterprise Search". **SEARCHER**. New Jersey. Vol 12; n. 10, pp. 8-17.

Battelle, John. 2005. **A Busca - Como o Google e Seus Competidores Reinventaram os Negócios e Estão Transformando Nossas Vidas**. Editora CAMPUS, São Paulo.

Berners-Lee, Tim, Hendler, James e Lassila, Ora. 2001. "The Semantic Web". **Scientific American**, 284(5), pp. 34-43.

Berners-Lee, Tim. 1997. **Realising the Full Potential of the Web**. Based on a talk presented at the W3C meeting, London,. http://www.w3.org/1998/02/Potential.html - Ultima visita em 12/07/2006.

Biever, Celeste. 2005. "Will Google help save the planet?" **New Scientist**. Vol. 187, nº 2512, pp. 28-29. 13 Aug.

Björneborn, Lennart, Ingwersen, Peter. 2004. Toward a basic framework for webometrics. **Journal of the American Society for Information Science and**

Technology, Vol. 55, nº 14, pp.1216-1227. Royal School of Library and Information Science. http://dx.doi.org/10.1002/asi.20077.

Brandt, Daniel. 2002. PageRank: Google's Original Sin. **NY Transfer News, Blythe Systems**. http://www.google-watch.org/pagerank.html, última visita em 09/09/2006.

Brin, Sergey and Page, Lawrence. 1998. **The anatomy of a large-scale hypertextual Web search engine**. In Proceedings of the Seventh international Conference on World Wide Web 7 (Brisbane, Australia). p. 107-117. DOI = http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

Byrne, Séamus. 2004. Stop Worrying and Learn to Love the Google-Bomb - School of Media and Communications. **Fibreculture Journal**, Issue: 3 http://www.journal.fibreculture.org/issue3/issue3_byrne.html - visitada em 17/09/2006.

Carvalho, L.A.V., 2002, **Datamining - A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**, Editora Érica, São Paulo, Segunda Edição.

Cho, Junghoo, Garcia-Molina, Hector and Page, Lawrence. 1998. **Efficient Crawling Through URL Ordering**. Seventh International Web Conference (WWW 98). Brisbane, Australia, April 14-18. http://dx.doi.org/10.1016/S0169-7552(98)00108-1.

Cho, Junghoo, Roy Sourashis e Adams, Robert E. 2005. **Page Quality: In Search of an Unbiased Web Ranking**. In Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data (Baltimore, Maryland, June 14 - 16, 2005). SIGMOD '05. ACM Press, New York, NY, 551-562. DOI= http://doi.acm.org/10.1145/1066157.1066220.

Couvering, Elizabeth Van .2004. **New Media? The political Economy of Internet Search Engines**, a paper presented to The Communication Technology Policy section for the 2004 Conference of the International Association of Media & Communication Researchers (IAMCR), Porto Alegre, Brazil, July 25-30.

Demo, Pedro. 2000. Ambivalência da sociedade da informação. **Ciência da informação***,* Brasília, v. 29, n. 2, pp. 37-42 maio/ago.

Dias, Alejandro M. **Through the Google Goggles: Sociopolitical Bias in Search Engine Design** – 2005. http://epl.scu.edu:16080/~stsvalues/readings/Diaz_thesis _ final.pdf . visitado em 26/11/2006.

Diaz-Isenrath, Cecilia. 2005. Um Estudo sobre Google: Questões para uma Leitura Micropolítica das Tecnologias da Informação. **Liinc em revista**, v.1, n.2, setembro 2005, pp.101-127. http://www.liinc.ufrj.br/revista 101.

Diligenti, M., Gori, M., and Maggini, M. 2002. **Web page scoring systems for horizontal and vertical search**. In Proceedings of the 11th international Conference on World Wide Web (Honolulu, Hawaii, USA, May 07 - 11, 2002). WWW '02. ACM Press, New York, NY, 508-516. DOI= http://doi.acm.org/10.1145/511446.511512.

Diligenti, Michelangelo, Gori, Marco, and Maggini, Marco. 2002. **Web page scoring systems for horizontal and vertical search**. In Proceedings of the 11th international Conference on World Wide Web (Honolulu, Hawaii, USA, May 07 - 11). WWW '02. ACM Press, New York, NY, 508-516. DOI= http://doi.acm.org/10.1145/511446.511512.

Dreilinger, Daniel and Howe, Adele E. 1997. **Experiences with selecting search engines using metasearch**. ACM Trans. Inf. Syst. 15, 3 (Jul. 1997), 195-222. DOI= http://doi.acm.org/10.1145/256163.256164.

Eastman, Caroline M. e Jansen, Bernard J. 2003. **Coverage, relevance, and ranking: The impact of query operators on Web search engine results**. ACM Transactions on Information Systems, Vol. 21, Is. 4, pp. 383-411. DOI= http://doi.acm.org/10.1145/944012.944015.

Eiron, Nadav, McCurley, Kevin S., and Tomlin, John A. 2004. **Ranking the web frontier**. In Proceedings of the 13th international Conference on World Wide Web (WWW '04). ACM Press, New York, NY, 309-318. DOI= http://doi.acm.org/10.1145/988672.988714.

Goldman, Eric. 2005. Deregulating Relevancy in Internet Trademark Law. **Emory Law Journal**, Vol. 54. Encontrado também em SSRN: http://ssrn.com/abstract=635803.

Goldman, Eric. 2006. Search Engine Bias and the Demise of Search Engine Utopianism. **Yale Journal of Law & Technology, Available at SSRN**: http://ssrn.com/abstract=893892.

GOOGLE Inc. Política de Privacidade do Google, **Centro de Privacidade do Google**, publicada em 14 de outubro de 2005. http://www.google.com/intl/pt-BR/privacy_faq.html, ultima visita em 04/01/2007.

Green, David C. 2003. Search Engine Marketing: Why it Benefits Us all. **Business Information Review**, Vol. 20, No. 4, 195-202. DOI: 10.1177/0266382103204005.

Gruhl, Daniel, Meredith, Daniel N., Pieper, Jan H., Cozzi, Alex and Dill, Stephen. 2006. **The web beyond popularity: a really simple system for web scale RSS**. In Proceedings of the 15th international Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 183-192. DOI= http://doi.acm.org/10.1145/1135777.1135809.

Haag, Carlos, 2006, **"**Quem tem poder sobre o quarto poder? Blogs desmitificam jornalismo e pautam noticiários"**. Revista Pesquisa FAPESP**, Edição Impressa 126 - Agosto 2006, pp. 81 - 85.

Hiler, John. **Google Time Bomb - Will Weblogs blow up the world's favorite search engine?** - http://www.microcontentnews.com/articles/googlebombs.htm - ultima visita em 07/04/2006.

Introna, L. D. & Nissenbaum, H. 2000. "Shaping the Web: why the politics of search engines matters". **The Information Society**, 16, 169-185.

Jeanneney, Jean-Noël. 2006. **Quando o Google Desafia a Europa: em defesa de uma Reação**. ContraCapa, Rio de Janeiro.

Kahn, Richard and Kellner, Douglas. 2004. New Media and Internet Activism: From the 'Battle of Seattle' to Blogging. **New Media Society**, Vol 6: 87 - 95. DOI: 10.1177/1461444804039908.

Kobayashi, M. Takeda, K. 2000. "Information Retrieval on the Web". **ACM Association For Computing Machinery**. Vol 32; Part 2, Pages 144-173.

Koster, Martijn. 1995. "Robots in the Web: threat or treat?" **ConneXions - The Interoperability Report**, Volume 9, nº 4, April.

Lawrence, Steve e Giles, C. Lee. 1999**. "**Accessibility of information on the web". **Nature**, Volume 400, p. 107, DOI = http://dx.doi.org/10.1038/21987.

Lawrence, Steve. 2001. "Free online availability substantially increases a paper's impact". **Nature**, Volume 411, Number 6837, p. 521. DOI: 10.1038/35079151.

Lawson, Stephen. 2006. "Google attempts to get lawsuit dismissed, Should be able to use any criteria it wants". **PCADVISOR**. 2006 July, Issue 137 - http://www.pcadvisor.co.uk/news/index.cfm?newsid=6509.

Lerner, Reuven. 2006. "At the forge: Google web services". **Linux Journal**. Volume 2006, Issue 145 (May. 2006).

Lévy, Pierre. 1996. **As Tecnologias da Inteligência. O futuro do Pensamento na Era da Informática. A Cultura da Informação e a Educação**. Editora 34. São Paulo.

_____. 2000. **Cibercultura**. 2ª Edição. Editora 34. São Paulo.

Mann, Charles C. 2006. "How Click Fraud Could Swallow the Internet". **Wired Magazine**, Issue 14.01 - Vol. 14 - January. http://www.wired.com/wired/archive/14.01/fraud.html. Visitada em 02/02/2007.

Marcondes, Christian A. 1998. **Programando em HTML 4.0**. Editora Erica Ltda. São Paulo.

Mayer, Marissa. 2005. Googlebombing 'failure'. - http://googleblog.blogspot.com/2005/09/googlebombing-failure.html – (Blog official da Google Inc.). Visitada em 02/04/2006.

Maze, Susan, Moxley, David e Smith, Donna. 1997. **Neal-Schuman authoritative guide to Web search engines**. New York: Neal-Schuman Plublishers, Inc.

Mukhopadhyay, Tridas, Rajan, Uday, and Telang, Rahul. 2004. **Competition between Internet Search Engines**. In Proceedings of the 37th Annual Hawaii international Conference on System Sciences (Hicss'04) - Track 8 - Volume 8 (January 05 - 08, 2004). HICSS. IEEE Computer Society, Washington, DC, 80216.1. http://portal.acm.org/citation.cfm?id=963176&coll=Portal&dl=GUIDE&CFID=432481&CFTOKEN=24292800#.

Nicholson, Scott, Sierra, Tito, Eseryel, U. Yeliz, Park, Ji-Hong, Barkow, Philip, Pozo, Erika J. e Ward, Jane. 2006. **How much of it is real? Analysis of paid placement in Web search engine results**, Journal of the American Society for Information Science and Technology, Vol. 75, n. 4 pp 448-461, University School of Information Studies, Syracuse, NY, DOI: 10.1002/asi.20318, http://dx.doi.org/10.1002/asi.20318.

Ninan, Sevanti. "Searching possibilities". **The Hindu - Online edition of India's National Newspaper** http://www.hindu.com/thehindu/2001/08/26/stories/13260696.htm. Ultima visita em 26/10/2005.

Nwana, Hyacinth S. 1996. "Software Agents: An Overview". **The Knowledge Engineering Review**, vol 11, nº 3, pp. 1-40.

Olsen, Stefanie. 2002. "Does search engine's power threaten Web's independence?" **CNET News.com**, http://news.com.com/2009-1023-963618.html, ultima vista em 24/01/2007.

Pasquale, Frank A. 2006. "Rankings, Reductionism, and Responsibility" (February 25, 2006). **Seton Hall Public Law Research Paper No. 888327** Available at SSRN: http://ssrn.com/abstract=888327.

Pelline, Jeff. 1998. “Pay-for-placement gets another shot”. **CNET News.com**. http://news.com.com/2100-1023-208309.html, ultima visita em 15/01/2007.

Schwartz, Candy. 1998. “Web search engines”. **Journal of the American Society for Information Science**. Vol: 49, nº: 11, pp: 973-982. New York.

Souza, Renato R. e Alvarenga, Lídia. “A Web Semântica e suas contribuições para a ciência da informação”. **Ci. Inf.**, Brasília, v. 33, n. 1, 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652004000100016&lng=pt&nrm=iso>. Acesso em: 27 Dez 2006. doi: 10.1590/S0100-19652004000100016.

Sullivan, Danny. 2001. “The Evolution Of Paid Inclusion”. **Search Engine Watch**. http://searchenginewatch.com/showPage.html?page=2163971, ultima visita em 22/01/2007.

Tatum, Clifford. 2005. “Deconstructing Google bombs: A breach of symbolic power or just a goofy prank?” **First Monday**, volume 10, number 10 (October 2005), http://firstmonday.org/issues/issue10_10/tatum/index.html - visitada em 14/09/2006.

United States Securities and Exchange Commission – SEC. Registration Statement of Google Inc. (form S-1 under the Securities Act.) April, 2004. Disponível em http://www.sec.gov/Archives/edgar/data/1288776/000119312504139655/ds1a.htm. Ultima visita em 22/02/2007.

United States Patent and Trademark Office (USPTO), 1998. “Method for node ranking in a linked database”. Appl. No.: 09/004,827 http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=6285999.PN.&OS=PN/6285999&RS=PN/6285999. Última visita em 18/02/2007.

Van Wel, Lita. and Royakkers, Lambèr. 2004. **Ethical issues in web data mining**. Ethics and Inf. Tech. 6, 2 (Jun. 2004), 129-140. DOI= http://dx.doi.org/10.1023/B:ETIN.0000047476.05 912.3d.

Vaughan, Liwen e Thelwall, Mike. 2004. “Search engine coverage bias: evidence and possible causes”. **Information Processing and Management: an International Journal**. N.40, pág 693-707. DOI: http://dx.doi.org/10.1016/S0306-4573(03)00063-3.

Vise, David A. e Malseed, Mark. 2007. **Google: a história do negócio de mídia e tecnologia de maior sucesso dos nossos tempos**. Editora Rocco, Rio de Janeiro.

Walker, Leslie. 2006. “Forgot What You Searched For? Google Didn't”. **The Washington Post Journal**. Saturday, January 21, 2006; Page D01.

Wooldridge, M. Jennings, N. R. 1995. “Intelligent agent: theory and practice”. **The Knowledge Engineering Review**, vol 10, nº 2, pag. 120.

Yang, Hsin-Chang and Lee, Chung-Hong. 2003. **A Text Mining Approach on Automatic Generation of Web Directories and Hierarchies**. In Proceedings of the IEEE/WIC international Conference on Web intelligence (October 13 - 17, 2003). WI. IEEE Computer Society, Washington, DC,625. http://doi.ieeecomputersociety.org/10.1109/WI.2003.124 1282.

Zhou, Wei; Smalheiser, Neil and Yu, Clement. 2006. “A tutorial on information retrieval: basic terms and concepts”**. Journal of Biomedical Discovery and Collaboration,** 1,2. DOI:10.1186/1747-5333-1-2.