Recuperação de informações na *Web*

Regina Meyer Branski

Pesquisadora do Instituto de Economia da Universidade Estadual de Campinas. Doutoranda em Engenharia da Produção - Escola Politécnica da Universidade de São Paulo. *e-mail*: branski@obelix.unicamp.br

Existem na Web milhares de páginas cobrindo os mais variados assuntos e interesses. Localizar estas informações não é tarefa simples. As ferramentas de busca são instrumentos fundamentais para auxiliar nesta tarefa. Pretende-se mostrar as diferenças nas formas de operação das diversas ferramentas de busca atualmente existentes na Web e como suas peculiaridades podem afetar os resultados de uma pesquisa. Conhecendo suas características e modo de funcionamento é possível extrair todo o potencial de cada ferramenta e localizar as informações desejadas mais eficientemente. Discute a Web oculta, grande volume de informação que não é indexado pelos mecanismos de busca tradicionais.

Palavras-chave: Internet; *Web;* Mecanismos de busca; Ferramentas de busca; Buscadores; Programas de busca; Catálogos; Diretórios; Metapesquisadores; *Web* oculta

Recebido em 22.09.2003 Aceito em 28.10.2003

Introdução

A recuperação de informações em banco de dados é um assunto bastante discutido pelos profissionais da ciência da informação. O advento da Internet tornou esta questão ainda mais premente. A difusão de seu uso ampliou a necessidade de acessar, de forma rápida e precisa, informações armazenadas em banco de dados gigantescos.

A Internet é um conjunto de inúmeras redes de computadores, conectadas entre si, que permite a comunicação, partilha de informações, programas e equipamentos entre seus usuários. Constitui a infra-estrutura sobre a qual trafegam grande volume de informações e outros serviços.

A Internet teve origem em uma rede, a Arpanet, criada pelo Departamento de Defesa dos Estados Unidos no inicio dos anos 70, interligando vários centros militares e de pesquisa com objetivos de defesa, na época da Guerra Fria. A tecnologia desenvolvida permitia a comunicação entre diferentes sistemas de computação, o que possibilitou a incorporação de outras redes experimentais que foram surgindo ao longo do tempo.

Atualmente, parte significativa da informação disponível na Internet é fornecida através da *World Wide Web* ou *Web*. A *Web* é um sistema baseado em hipertexto, que constitui a capacidade de ligar palavras ou frases de uma página *Web* a outros recursos da Internet através de *links*. Quando se clica com o *mouse* sobre um *link,* ele remete para outro ponto dentro do mesmo documento, para outra página *Web* ou mesmo para outro *site* diferente daquele originalmente acessado. Pode-se, ainda, abrir automaticamente uma mensagem de *e-mail*, abaixar algum *software* ou artigo etc.

Estão disponíveis na *Web* milhares de páginas cobrindo os mais variados assuntos e interesses. Estimativas recentes afirmam existir cerca de 2,5 bilhões de documentos, com uma taxa de crescimento de 7,5 milhões ao dia[1]. Mas, diferentemente das bibliotecas, os documentos da Internet não estão classificados segundo um padrão determinado. Portanto, o usuário precisa localizar informações de um grande volume de páginas disponíveis, sem qualquer organização.

Encontrar a informação depende, principalmente, do uso eficiente das ferramentas de busca. Para explorar todo o potencial dos buscadores, o usuário precisa conhecer: como é coletada e estruturada a informação em diferentes bancos de dados; suas características e limitações; todas as possíveis formas de interação e suas linguagens de busca.

Conceitua-se estrutura da informação como a sua organização lógica para posterior recuperação e linguagem de busca como os comandos que permitem a recuperação da informação através de palavras contidas nos títulos, resumos ou outros campos de dados.

Este artigo pretende mostrar as diferentes formas de coletar e estruturar as informações que caracterizam os diversos buscadores disponíveis na Internet, como elas afetam os resultados das pesquisas e a importância do uso da linguagem controlada no acesso a estes bancos de dados. Este conhecimento capacitará o usuário a recuperar, de forma rápida e precisa, a informação que precisa.

[1] "*Sizing the Internet*" (http://www.cyveillance.com/web/us/downloadas/Sizing_the_Internet.pdf)

O que são buscadores e como funcionam

Buscadores, ferramentas de busca ou mecanismos de busca são sistemas especializados utiliazados na recuperação de informações na Internet. Eles coletam informações sobre os conteúdos dos *sites* e os armazenam em bancos de dados que estão disponíveis para consulta. Realizando uma busca, o usuário poderá descobrir a localização exata das informações que deseja.

As informações são armazenadas em bancos de dados porque são flexíveis, fáceis de operar e manter. O acesso a estes bancos de dados em um ambiente *Web* é possível graças a uma interface especial, capaz de traduzir os dados armazenados para uma linguagem compreendida pelo *Netscape, Microsoft Explorer* ou outro navegador que estiver sendo utilizado. Entretanto, para que o usuário acesse o conteúdo que está por trás da *Web,* ele precisa visitar a página de interface e realizar uma pesquisa específica.

O usuário digita alguma expressão, geralmente uma palavra ou frase no campo de busca, e, em seguida, solicita a pesquisa. Os buscadores procuram a ocorrência da linguagem de busca nas informações armazenadas em seus bancos de dados. Ou seja, quando se realiza uma busca não se está pesquisando diretamente na Internet, mas no banco de dados do buscador escolhido.

As ferramentas de busca oferecem, como resposta ao usuário, páginas onde estão relacionados todos os *sites* armazenados em seu banco de dados onde foram verificadas a ocorrência da linguagem de busca. Os resultados são apresentados na forma de *links* de hipertextos, isto é, clicando com o *mouse* sobre uma das frases realçadas (*links)*, o próprio *site*, que está fora do banco de dados do buscador, é trazido para o computador do usuário.

As páginas que os mecanismos de busca oferecem como resposta não são armazenadas no servidor. Elas simplesmente desaparecem após a consulta. Estas páginas são transitórias porque é mais barato e fácil reconstruí-las novamente do que armazená-las com todas as possíveis opções existentes de consulta.

A eficiência de um buscador será avaliada pela sua capacidade em apresentar, logo nas primeiras linhas, informações que atendam às necessidades dos usuários. Para isso, seus organizadores construiram um banco de dados amplo e com informações de qualidade. Eles devem, ainda, ser capazes de entender o que o usuário - a maioria das vezes inexperiente - quer e recuperar as informações adequadas.

A eficiência do usuário, por sua vez, depende de sua capacidade em oferecer ao banco de dados elementos suficientes para que sejam selecionados, a partir da totalidade das informações armazenadas, um conjunto de itens que constituam a resposta que procura.

As definições, por parte dos desenvolvedores dos *sites* que serão recuperados através do uso da base de dados e a forma como serão ordenados contribui decisivamente para a eficiência do buscador. A maioria deles utiliza critérios que envolvem localização e freqüência. Analisando os títulos, resumos e a freqüência da ocorrência da linguagem de busca nos documentos que compõem sua base de dados, definem os *sites* e a ordem em que serão apresentados aos usuários.

O mecanismo de busca *Google*[2] inovou no critério utilizado para apresentação dos documentos recuperados, alcançando resultados bastante satisfatórios. Este buscador define seus resultados de acordo com o número de *links* apontando para cada um dos documentos armazenados em sua base de dados. Isto é, na relação dos documentos recuperados pelo buscador, ocuparão os primeiros lugares os *sites* que tiverem sidos mais citados por outros *sites*. Esta forma de estruturar a informação tem como premissa a idéia de que os *sites* mais populares oferecem informações de melhor qualidade.

Finalmente, os conteúdos armazenados nestes buscadores constituem apenas parte das informações disponíveis na Internet. Cada ferramenta de busca tem armazenada em sua base de dados um subconjunto particular de *sites* selecionados. A forma utilizada pelo buscador para coletar as informações que formarão este subconjunto tem impacto direto nos resultados que o usuário obterá. Pode-se identificar duas categorias de buscadores: diretórios por assunto e programas de busca. Serão discutidos cada um deles a seguir.

a) Catálogos ou diretórios por assunto

Os catálogos ou diretórios por assunto precederam os programas de busca e constituíram a primeira tentativa de estruturar e recuperar recursos na *Web*. Foram criados quando a quantidade de recursos disponíveis ainda permitia a coleta das informações manualmente.

Nos diretórios por assunto, as informações que compõe o banco de dados são coletadas de duas formas:

•através de busca realizada por seus editores, que visitam inúmeros *site*s e incluem os de interesse no banco de dados, acompanhados de uma breve descrição de seus conteúdos;

• através de solicitação de inclusão enviada pelo autor interessado em ter seu *site* catalogado. O autor envia uma breve descrição do conteúdo, e os editores visitam o *site*, aceitando ou não sua inclusão.

As informações são organizadas e classificadas hierarquicamente em categorias temáticas pelos editores. Parte-se das categorias mais amplas para as mais específicas. Por exemplo, no popular *Yahoo*![3], informações sobre tubarões estão classificadas na categoria *Ciência » Animais, insetos e bichos de estimação » Vida aquática » Peixes » Espécies » Tubarões.*

Pode-se consultar um diretório digitando uma palavra ou frase no campo de busca ou explorando suas categorias. O catálogo verificará a ocorrência da linguagem de busca no título e na descrição enviada pelo autor ou compilada pelos editores, não sendo considerado o texto integral do *site*.

Os catálogos podem ser bastante úteis quando o número de respostas obtidos nos programas de busca for excessivo ou, ainda, quando a informação desejada estiver contemplada em uma das categorias disponíveis. Por exemplo, endereço na Internet de todas as universidades americanas. Os editores certamente já coletaram estas informações e, melhor ainda, organizaram e classificaram todas elas, tornando a tarefa do usuário bem mais simples.

[2] http://www.google.com

[3] http://www.yahoo.com

Os bancos de dados dos catálogos são menores e menos atualizados que os dos programas de busca. Entretanto, as informações são mais selecionadas por passarem pelo crivo dos editores.

Entre os diretórios mais populares destacam-se o *Yahoo!* (http://www.yahoo.com), *Lycos* (http://www.lycos.com), *HotBot* (http://www.hotbot.com) e, no Brasil, o Cadê? (http://www.cade.com.br), Achei (http://www.ahei.com.br) e *Yahoo!* Brasil (http://www.br.yahoo.com).

b) Programas de busca

Os programas de busca surgiram quando a quantidade de informações disponíveis na *Web* atingiram proporções que dificultavam sua coleta manual. Estas ferramentas criam seus bancos de dados automaticamente utilizando *softwares* conhecidos como *spiders* ou robôs.

Os robôs percorrem a rede coletando informações. Iniciam seu caminho a partir de um conjunto de páginas selecionadas por seus administradores que são escolhidas por serem populares, de alta qualidade ou por conterem grande quantidade de *links*. Os robôs visitam os *sites*, lendo seu conteúdo, armazenando cada uma das palavras encontradas e seguindo seus *links* para outras páginas[4]. Seguem os *links* para encontrar outros links e ir, sucessivamente, adicionando os endereços ao banco de dados. Diferentemente dos catálogos, os *sites* são incluídos no banco de dados sem nenhuma classificação ou descrição de seu conteúdo.

Os programas de busca podem coletar as informações de diferentes formas. Alguns incluem em seus bancos de dados o conteúdo integral dos sites, lendo e registrando cada palavra do início ao final do documento. Outros, limitam-se a registrar o título e um resumo construído automaticamente pelo robô. Outros ainda, o título e as primeiras linhas do *site*.

Os robôs voltam aos *sites* em intervalos regulares para verificar alterações e manter o sistema atualizado. Neste processo eliminam páginas desativadas, incluem novas e incorporam as modificações.

Os programas de busca são mais abrangentes que os catálogos uma vez que os robôs registram toda a informação encontrada. Fornecem, portanto, uma visão mais ampla do conteúdo disponível na Internet sobre um assunto. Por outro lado, por terem bancos de dados com milhões de páginas, a busca pode ser mais imprecisa, retornando um grande número de respostas insatisfatórias.

Entre os programas mais conhecidos destacam-se *Alta Vista* (http://www.altavista.com), *Northern Light* (http://www.northernlight.com), Google (http://www.google.com) e, no Brasil, TodoBr (http://www.todobr.com.br) e Radar UOL (http://www.radaruol.com.br).

O leitor deve deduzir, a partir do exposto, que o resultado de uma mesma pesquisa utilizando um diretório e um programa de busca serão diferentes. Haverá também discrepâncias quando forem utilizados dois programas ou dois catálogos. Estas diferenças decorrem não só da forma de construção dos bancos de dados, como também do modo de funcionamento de cada ferramenta.

Entre os catálogos, a variação de resultados pode ser atribuída: aos *sites* selecionados que comporão o banco de dados; aos critérios utilizados para classificação das informações.

[4] Se uma página não tem nenhum *link* apontando para ela, os robôs não tem como encontrá-la. Neste caso, a única maneira de incluí-la em um programa de busca é enviando um pedido de inclusão. Todas as ferramentas de busca oferecem esta possibilidade.

Entre os programas de busca, a variação dos resultados pode decorrer: da definição das páginas iniciais, a partir das quais o robô percorrerá a rede para coletar informações e alimentar seu banco de dados, da forma como registram as informações de cada *site* (se armazenam o texto integral, somente o título e um pequeno resumo construído de forma automática, o título e as primeiras linhas do *site* etc.).

Assim, a utilização de mais de uma ferramenta garante uma maior cobertura e, possivelmente, um resultado mais satisfatório. Deve-se considerar ainda que nenhum dos buscadores incluem em seus bancos de dados a totalidade dos *sites* existentes na Internet[5]. Operam em suas próprias bases compostas de *sites*, textos e descrições selecionados a partir da totalidade dos documentos da rede. Pesquisa-se, portanto, em um subconjunto relativamente pequeno dos sites que compõem a *World Wide Web*.

c) Refinando a pesquisa

Quem ainda não se defrontou com a seguinte situação: o resultado de uma busca, especialmente quando são utilizados programas de busca, apresenta um número excessivo de respostas, sendo que a maioria delas não tem relação com o que se procura. O passo seguinte é conhecido. Gasta-se um tempo enorme para se selecionar o que é relevante disponibilizado pelos sites.

Um modo de contornar este problema é através do uso de refinamentos, que são comandos que permitem limitar e controlar a ação das ferramentas de busca. Através deles é possível definir melhor o objeto de interesse e tornar a pesquisa mais eficiente.

Quando se digita no campo de busca a expressão comércio exterior, por exemplo, a ferramenta localizará em seu banco de dados páginas que contenham as duas palavras, mesmo que distantes uma da outra. Se se colocar a expressão entre aspas, pode-se obter documentos onde tais palavras aparecem exatamente nesta ordem. Através dos refinamentos pode-se localizar arquivos de imagem, limitar a busca a um certo domínio ou título, identificar palavras com diferentes terminações etc.[6]

Combinando diversas formas de refinamentos, pode-se controlar ainda mais o campo de ação dos buscadores. Digitando no programa de busca Alta Vista[7] a expressão *title:"comércio internacional" url:edu,* pode-se obter *sites* com exatamente este título provenientes exclusivamente de instituições de ensino.

O controle da linguagem de busca melhora significativamente a eficiência da pesquisa. Mas, cada buscador tem suas próprias características, aceitando ou não determinados refinamentos. Assim, é importante o conhecimento das particularidades de cada ferramenta para ser capaz de extrair todo o seu potencial.

No Quadro 1 estão relacionados alguns refinamentos e suas funções. No Quadro 2 estão descritos os refinamentos aceitos por alguns buscadores selecionados. Foram analisados os programas de busca estrangeiros *Alta Vista* (http://www.altavista.com), *Google* (http://www.google.com), *Northern Light* (http://www.northernlight.com) e os brasileiros Radar UOL (http://www.radaruol.com.br) e TodoBr (http://www.todobr.com.br). Dentre os catálogos mereceram especial atenção *Yahoo!* (http://www.yahoo.com), *HotBot* (http://www.hotbot.com) e Lycos (http://www.lycos.com) e os brasileiros Cadê? (http://www.cade.com.br), *Yahoo*! Br (http://br.yahoo.com) e Aonde (http://www.aonde.com.br).

[5] Segundo o *site Search Engine Watch (http://www.searchenginewatch.com/reports/sizes.html)*, o programa de busca *Google* tem a maior base de dados, indexando cerca de 75% das páginas existentes na Web, sendo seguido pelo *Fast (http://www.alltheweb.com)* com com 31% e pelo Alta Vista com 27,5%.

[6] Os mecanismos de busca adotam automaticamente o sinal + entre as palavras. Isto significa que na ausência de comandos entre os termos de busca, a ferramenta retornará os documentos onde estão presentes *todas* as palavras digitadas, sem contudo observar a ordem de apresentação.

[7] *http://www.altavista.com*

QUADRO 1 - Refinamento e suas funções

Objetivo	Comandos	Função	Exemplos
	nenhum	Localiza páginas que contenham todos os termos pesquisados, qualquer que seja a ordem em que são apresentados	comércio exterior retornará documentos onde constem as duas palavras, qualquer que seja a ordem
Incluir	+ and e	Localiza páginas que contenham todos os termos pesquisados, qualquer que seja a ordem em que são apresentados	+comércio +trigo comércio AND trigo retornará documentos onde constem as duas palavras, qualquer que seja a ordem
Excluir	+ - not and not	Exclui páginas que contenham o termo selecionado	+planeta–casseta planeta NOT casseta planeta AND NOT casseta retornará páginas que contenham a primeira palavra, excluindo as que contenham a segunda.
Frase Exata	aspas	Localiza páginas que contenham a frase exata	"comércio internacional" retornará documentos onde constem as duas palavras na ordem exata
or ou	Buscar qualquer termo	Localiza páginas que contenham qualquer um dos termos pesquisados	feminina OR feminilidade retornará documentos onde constem qualquer uma das duas palavras
Proximidade	near	Define quão próximo os termos devem aparecer	lua NEAR rio retornará páginas que contenham as duas palavras separadas por uma certa distância, definida pelo pesquisador
Buscar *site*s	host:*site*:domain:	Localiza páginas num computador específico	host:nasa.gov venus localiza, somente no *site* da NASA, páginas que contenham a palavra Vênus"mars exploration" +domain:edulocaliza páginas sobre exploração de Marte provenientes, exclusivamente, de *site*s educacionais americanos
Buscar URL[8]	url:u:allinurl:inurl:	Localiza páginas que possuam uma palavra ou frase específica na URL	url:jardimlocaliza páginas que contenham em sua URL a palavra jardim
Buscar *links*	*link*:	Localiza páginas que possuam *links* apontando para uma página ou domínio particular	*link*:www.unicamp.br localiza páginas com *links* apontando para o *site* da Unicamp
Buscar no título	title:allintitle:intitle:t:	Localiza páginas que possuam uma palavra ou frase específica no título	title:"comércio internacional"localiza páginas cujo título contenha a frase comércio internacional
Terminação	*	Localiza palavras com diferentes terminações	femini* localiza páginas que contenham as palavras feminina, femininas, feminino, feminilidade, etc.

[8]Cada página de um *site* tem um endereço único denominado Uniform Resource Locator (URL) que possibilita sua localização por computadores no mundo todo. Por exemplo, a URL *http://www.eco.unicamp.br* identifica a página inicial do *site* do Instituto de Economia da Unicamp

QUADRO 2 - Refinamentos aceitos por buscadores selecionados

Objetivos		Comandos	Mecanismos de Busca										
			Programas de busca					Catálogos ou diretórios					
			Alta Vista	Google	Northern Light	Radar Uol	Todo Br[9]	Yahoo[10]	Lycos	HotBot	Cadê	Achei!	Yahoo Br
Incluir		nenhum	x	x	x	x	x	x	x	x	x	x	x
Incluir		+	x	x	x	x		x	x	x	x		x
Excluir		-	x	x	x	x		x	x	x	x		x
Boolean	Buscar qualquer termo	or	11		x		11		11	11	x	11	11
	Incluir	and			x		11				x		11
	Excluir	not			x				11	11			
		and not											
	Proximidade	near											
Frase Exata		aspas	x	x	x	x	x	x	x	x	x		x
Titulo		title:	x		x				x	x			
		t:						x			x		x
		allintitle:		x		x							
		intitle:		x		x							
Buscar *site*		host:	x						11				
		site:		x		x							
		domain:								x			
Buscar URL		url:	x										
		u:						x			x		x
		allinurl:		x		x							
		inurl:		x		x			11				
Buscar *links*		*link*:	x	x		x							
Terminação		*	x		x			x		x	x		x

d) Pesquisando na *Web*: estratégias, análise do assunto e escolha da ferramenta de pesquisa

Elaborar uma estratégia de busca é formular uma tática para recuperar informações armazenadas em um banco de dados. Uma estratégia de busca será eficiente se as informações recuperadas atenderem às necessidades do usuário.

Oldroyd e Citroen (1977) afirmam que, para planejar a estratégia de busca, o usuário deve decidir qual é a melhor base de dados para o seu tema, selecionar os termos de busca adequados e formular a estratégia.

Quando se utiliza um buscador, siogniofica que se está pesquisando no banco de dados daquela ferramenta. A maneira como o buscador coleta as informações que comporão sua base de dados e a forma como estrutura e recupera estas informações têm implicações importantes nos resultados que o usuário obterá. Portanto, para definir o banco de dados mais adequado aos seus objetivos, o usuário deve conhecer as características das diversas ferramentas disponíveis.

A familiaridade com a forma de funcionamento dos diferentes buscadores é um fator necessário mas não suficiente para garantir uma busca eficiente. O usuário deve, também, oferecer os elementos necessários para que sejam selecionados, a partir da totalidade das informações armazenadas no banco de dados, um conjunto de itens que constituam a resposta desejada.

Apesar do contínuo esforço dos desenvolvedores das ferramentas de busca para a criação de sistemas de recuperação amigáveis, com orientação através de *menus* ou oferecendo recursos especiais para usuários inexperientes, o processo de busca constitui, ainda, uma questão complexa. O usuário deve ser capaz de elaborar a linguagem de busca e a estratégia adequada. Para

[9] Programa de busca cujo banco de dados é composto somente de *sites* brasileiros (domínio br)

[10] Quando o termo de busca não é localizada no banco de dados do Yahoo!, a pesquisa é remetida automaticamente para o programa de busca Google.

[11] Opção disponível somente no modo avançado

isso, deve avaliar o que conhece a respeito do tópico e o que pretende saber e, a partir daí, executar os seguintes passos:

- formular a questão da pesquisa e sua abrangência
- identificar os conceitos importantes dentro da questão
- definir a linguagem de busca que identifique estes conceitos
- considerar sinônimos ou variações da linguagem
- preparar a lógica da busca (refinamentos).

O QUAD. 3, elaborado pelo *UC Berkeley Library*, indica algumas relações entre os objetivos da pesquisa e lógica de busca. À direita estão relacionados os objetivos e, à esquerda, os refinamentos adequados àquela necessidade.

QUADRO 3 – Objetivos da pesquisa X Lógica de busca

Procurando um nome próprio ou uma frase exata? 1. Nome de uma organização, sociedade ou movimento 2. Nome próprio ou um indivíduo 3. Seqüência de palavras precisas associadas, geralmente, a um assunto **É possível pensar em uma organização, nome próprio ou frase para pesquisar? Passo inicial para localizar o que se procura.**	**Pesquisador deve ser capaz de localizar frases. Exige que todos os termos apareçam na ordem exata em que foram digitados. Colocar a frase entre aspas. Exemplo:** • "*world health organization*" • "regina meyer branski" • "comércio internacional"
Os termos procurados são palavras comuns com muitos significados e vários contextos? 1. criança associada a televisão e violência **Lembre-se que na ausência de qualquer comando, os mecanismos de busca utilizam automaticamente o sinal +.**	**Utilizar AND pode ajudar** • criança and televisão and violência • criança televisão violência **Ou utilizar outros termos que possam levar ao mesmo assunto:** • jornalismo ética censura **Nova pesquisa ou controle dos termos resultantes podem ser úteis** • Após a pesquisa, submeter o resultado a outros aspectos
O resultado apresenta inúmeros termos que você não quer? • Pesquisa por engenharia biomédica e câncer traz inúmeros programas acadêmicos e o que se procura são artigos.	**Utilizar AND NOT pode ajudar** • "engenharia biomédica" AND câncer AND NOT "departamento de" AND NOT "escola de" **Ou seu equivalente** • +"engenharia biomédica" +câncer -"departamento de" -"escola de"
Existem sinônimos, variações de ortografia ou palavras estrangeiras para representar o que se está buscando? • Women, females com networking • Sarajevo, Sarayevo com peace • Literature, litterature com French, francaise	**Utilizar OR ou seu equivalente** • (Women OR female) AND networking • (Sarajevo OR Sarayevo) AND peace (literature OR litterature) AND (French OR francaise)
A pesquisa refere-se a *Home Pages* e/ou outros documentos básicos sobre os termos? • *Home Page* da American Dietetic Association • Páginas básicas sobre comércio exterior **A pesquisa refere-se a termos com vários finais possíveis?**	**Limitar a pesquisa ao título** • Title: "American Dietetic Association" • Title: "comércio exterior" **É possível adaptar todas as variações em um único termo de**

e) Estratégias não recomendadas

Algumas estratégias de busca são pouco eficientes e, portanto, devem ser evitadas:

· Exploração de catálogos. Recuperar documentos tentando combinar o assunto pesquisado com a categoria mais geral de uma hierarquia de assuntos. A partir daí, o usuário escolhe subcategorias que possam levá-lo ao objetivo pretendido. A principal dificuldade consiste em determinar sob qual categoria o assunto está classificado. Corre-se o risco de, após inúmeras tentativas, descobrir-se que o assunto procurado não está sob aquela

classificação. A categoria *saúde*, por exemplo, pode conter documentos sobre medicina, homeopatia, psiquiatria e esporte em determinado catálogo. Em outro catálogo, *medicina* pode incluir saúde, saúde mental e medicina alternativa e esporte pode estar classificado na categoria *estilo de vida*.

· Palavras-chave simples em bancos de dados amplos, como os programas de busca. Pesquisar com palavras-chave simples é buscar uma ou mais palavras, separadas por espaços, nas ferramentas de busca. Desta forma pode-se recuperar todos os endereços do banco de dados que contenham a palavra ou palavras pesquisadas. Em banco de dados extensos tal procedimento gera excesso de documentos sendo que, grande parte deles não têm são relevantes. Neste caso é aconselhável utilizar técnicas mais avançadas de pesquisa, controlando a linguagem de busca. Nos bancos de dados menores e em catálogos por assunto, entretanto, pesquisas utilizando palavras chaves simples podem fornecer uma boa aproximação.

Dessa forma, conclui-se a apresentação dos conceitos básicos dos mecanismos de busca: o que são, como funcionam, diferenças existentes na construção das bases de dados e controle da linguagem. As ferramentas de busca que serão apresentadas a seguir, embora tragam algumas novidades, podem ser enquadradas em uma das duas categorias já descritas: catálogos por assunto ou programas de busca.

Outras formas de localizar informações na *Web*

a) Mecanismos de busca especializados ou temáticos

Diferentemente dos mecanismos de busca genéricos, que armazenam informações sobre qualquer assunto, os buscadores especializados ou temáticos restringem-se a documentos de um campo específico. Suas bases de dados são compostas de informações pertencentes a uma única categoria como, por exemplo, comércio exterior, computação, medicina etc.

O número de buscadores especializados na Internet vem aumentando diariamente e cobrem quase todos os assuntos. Cada um deles tem conteúdo e abordagem únicos. A vantagem destas ferramentas sobre pesquisadores genéricos consiste em que, por serem especializados, apresentam resultados mais relevantes, num tempo de pesquisa menor.

Como os genéricos, estas ferramentas especializadas podem compilar seus bancos de dados através de robôs (como os programas de busca) ou utilizando editores que classificam os sites em tópicos (como os catálogos).

No caso dos programas de busca especializados, seu criador seleciona sites voltados para um assunto específico e o robô percorre os *links* a partir destes sites. Por exemplo, para elaboração de um banco de dados especializado em decoração, selecionam-se sites de qualidade sobre o assunto decoração e o robô percorre os *links* indicados por estes *sites*. Somente aqueles selecionados e as páginas indicadas farão parte do banco de dados do programa de busca. Veja, por exemplo, o *MedHunt* (*http://www.hon.ch/MedHunt*) cuja base de dados é criada a partir de *sites* selecionados da área médica ou,

ainda, o MP3.com (http://www.mp3.com) especializado na localização de músicas no formato MP3.

Nos catálogos especializados o banco de dados é compilado da mesma forma que nos catálogos genéricos. Os editores buscam novos *sites* e revisam as submissões apresentadas, classificando as informações em categorias.

Um catálogo especializado terá, provavelmente, mais *sites* sobre o assunto subdivididos em um número maior de categorias que os genéricos. Um exemplo é o catálogo *Advertising World* (http://advertising.utexas.edu/world), especializado em marketing ou o *Global Edge* (http://globaledge.msu.edu/ibrd/ibrd.asp), que coleciona *sites* voltados para o desenvolvimento de negócios internacionais.

Podemos localizar buscadores especializados digitando-se, no campo de busca a expressão "*specialized search engines*" ou o assunto de interesse acompanhado de uma das seguintes expressões: "*subject guides*", "*subject directories*", "*web directories*". Existem, ainda, catálogos que compilam, exclusivamente, buscadores especializados. Coletam programas de busca e catálogos especializados e os classificam em categorias. Veja, por exemplo:

· *http://www.internets.com*, catálogo onde os mecanismos de busca especializados estão classificados em 43 categorias;

· *SearchIQ* (http://www.zdnet.com/searchiq/subjects), guia de mecanismos de busca especializados organizado em 25 categorias e várias subcategorias,

· *Fossik.com* (http://www.fossick.com) lista mais de 3 mil pesquisadores especializados classificados em nove categorias e aproximadamente 50 subcategorias.

Finalmente, há na *Web sites* que oferecem uma coleção de *links* sobre um assunto específico. Pode-se dizer que estas coleções deram origem aos buscadores especializados e que são sua versão reduzida, características que não diminuem sua importância. São, geralmente, compilados por especialistas e seu conteúdo é selecionado cuidadosamente obedecendo critérios objetivos.

Bons exemplos podem ser encontrados em *http://lib.itg.be/biblinks.htm* que oferece *links* selecionados para área de saúde, no *http://www.soemadison.wisc.edu/ccbc/hplinks.htm* que inclui links selecionados sobre o personagem Harry Porter ou, ainda, no endereço *http://www.ruralllinks.com.br* que coleta informações na *Web* sobre agronegócios. Para localizar estes *sites* deve-se digitar o assunto de interesse e uma das seguintes expressões: *links, selected links, bookmarks, webligraphies* ou índices (ou seus correspondentes em português).

b) Bibliotecas virtuais

As bibliotecas virtuais existem exclusivamente na Internet. Mantidas geralmente por bibliotecas de universidades possibilitam o acesso a jornais, periódicos, livros e outras publicações que são digitalizados e disponibilizados na *Web*. Oferecem ainda, de modo geral, uma coleção de recursos da Internet que são coletados e organizados por pessoas qualificadas.

Esta coleção de *links* não pretende ser uma ampla lista de todos os *sites* de cada categoria, como os catálogos genéricos, mas uma seleção dos melhores. As fontes são selecionadas de acordo com a facilidade de uso,

qualidade, quantidade e origem das informações e freqüência das atualizações. Dentre as bibliotecas virtuais, destacam-se:

- *Internet Public Library* (http://www.ipl.org) que mantém uma coleção de mais de 40 mil recursos da Internet selecionados, organizados e descritos por bibliotecários;
- *Scout Report Archives* (http://scout.cs.wisc.edu/archives/), mantido por educadores e bibliotecas da Universidade de Wisconsin, oferece mais de 10 mil sites de valor educacional;
- *Infomine* (http://infomine.ucr.edu), mantido por diversas universidades americanas, oferece uma coleção com cerca de 23 mil recursos educacionais, entre eles bancos de dados, jornais, revistas eletrônicas, artigos etc. e
- *Britannica* (http://www.britannica.com), mantido pela Enciclopédia Britannica, coleta e classifica os melhores sites em diversas áreas, além de oferecer acesso online à enciclopédia. Os links são classificados e acompanhados de um breve sumário do conteúdo.

c) Mecanismos de metabusca ou metapesquisadores

Nos buscadores tradicionais submete-se os termos de busca a um único banco de dados e recebe-se uma relação dos documentos onde constam os termos pesquisados. Utilizando metapesquisadores o usuário está buscando, simultaneamente, em vários buscadores.

Os metapesquisadores não possuem banco de dados próprio e funcionam como um agente intermediário que repassa a pesquisa, obtém as respostas dos buscadores individualmente e, então, apresenta um resultado unificado, extraído das diversas fontes. Em poucos segundos os metapesquisadores compilam e apresentam os resultados obtidos em diversos mecanismos de busca.

Embora o seu uso possa significar economia de tempo, já que a pesquisa é feita em um único *site*, a qualidade dos resultados varia muito de acordo com a ferramenta escolhida.

As deficiências decorrem, principalmente:

- da forma como apresentam os resultados.

O ideal é que as respostas obtidas sejam integradas, ordenadas por relevância e os resultados duplicados sejam eliminados. Nem todos os mecanismos de metabusca trabalham desta forma. Alguns agrupam os resultados e os mostram em seqüência, dificultando a análise das ferramentas individualmente.

- da incapacidade de manipular pesquisas complexas.

O ideal é que as pesquisas sejam formatadas de acordo com os refinamentos aceitos por cada ferramenta de busca individualmente: quando um buscador é submetido a refinamentos que não processa, ocorrem erros e resultados inadequados. Sendo assim, a utilização de metapesquisadores é mais eficiente quando as pesquisas são simples.

Os metapesquisadores podem ser úteis quando se deseja obter um número pequeno de resultados relevantes, localizar tópicos pouco explorados

ou ter uma visão geral dos documentos disponíveis na *Web* sobre determinado assunto. Entretanto, retorna um número limitado de resultados que não representam o todo, oferecendo uma visão superficial, e muitas vezes distorcida, das bases de dados dos buscadores analisados. Portanto, seu uso não elimina a necessidade de uma busca individual nos diversos mecanismos de busca para uma boa estratégia de pesquisa.

Dentre os metapesquisadores destacam-se:

- *Ixquick* (http://ixquick.com)

Busca em inúmeros índices, catálogos, jornais e multimídia. Apresenta os dez primeiros resultados de cada mecanismo de busca eliminando as duplicações. É capaz de trabalhar com pesquisas complexas.

- *Profusion* (http://www.profusion.com)

Busca em nove mecanismos de busca: *Altavista, Yahoo!, Infoseek, LookSmart, Excite, Magellan, WebCrawler, GoTo* e *Google.* Permite a organização dos resultados por vários critérios. Seu grande diferencial está em formatar as perguntas de acordo com a sintaxe aceita por cada um dos serviços de busca, individualmente. Os resultados finais são ordenados e as entradas duplicadas são removidas tornando o resultado final mais fácil de se analisar.

- *Metaminer* (http://miner.bol.com.br)

Metapesquisador brasileiro. Busca nos pesquisadores Achei e Radar UOL e nos estrangeiros AOL e *Looksmart*.

d) *Web* Oculta

Web oculta é uma parte importante da *Web* na qual os mecanismos de busca tradicionais não podem ou não querem incluir em seus bancos de dados. Sendo assim, estes *sites* não aparecem nos resultados apresentados por estas ferramentas de busca. Estima-se que esta parte oculta da *Web* tenha mais que o dobro do tamanho da parte visível e seu conteúdo é bastante relevante.

Há, basicamente, duas razões para estes *sites* estarem fora dos bancos de dados de grande parte dos buscadores:

- questões técnicas que impedem o acesso dos *spiders* a alguns tipos de *site*s.
- por decisão dos administradores dos mecanismos de busca.

Serão discutidas cada uma delas detalhadamente.

e) Questões técnicas

Os *softwares* conhecidos como robôs ou *spiders*, constróem seus bancos de dados automaticamente. A partir de uma relação de páginas selecionadas, seguem todos os *links* encontrados para armazenar as informações e alimentar seus bancos de dados. Estes robôs não são capazes de digitar informações ou definir opções. Portanto, não podem incluir em seus bancos de dados *site*s que exijam tais tipos de comandos.

A forma de operar dos robôs, e suas limitações, provocam a exclusão dos seguintes tipos de *site*s:

· *site*s desconectados. Para que um mecanismo de busca indexe uma página, o o autor envia um pedido de submissão ou o robô descobre a página por si próprio, encontrando um *link* a partir de uma página conhecida. Páginas *Web* que não forem diretamente submetidas ao mecanismo de busca e não tenham *links* apontando para elas estão desconectadas e, portanto, fora do alcance dos robôs.

· *site*s que exijam que se digite alguma informação para serem acessados. Incluem-se neste caso as páginas que requerem registro do usuário. Assim, para acessar o conteúdo é preciso digitar a senha e o *login*. Os administradores exigem estas informações para controlar o acesso a *sites* de uso restrito.

· *site*s que funcionam como interface para outros bancos de dados requerem do usuário a definição de uma série de opções para editar o conteúdo que será acessado. Esta exigência impede os robôs de incluir, em seus bancos de dados, outros bancos de dados.

Os mecanismos de busca genéricos não são capazes de acessar os conteúdos das páginas transitórias geradas por outros bancos de dados. Quando um *spider* se depara com um banco de dados isto funciona como se encontrasse uma biblioteca com portas de segurança invioláveis. São capazes de lembrar o endereço da biblioteca mas não podem dizer nada sobre os livros, revistas ou outros documentos armazenados.

Os robôs não tem dificuldade em encontrar a interface de um banco de dados porque se assemelham a outras páginas *Web* que utilizam formas interativas. Mas, os comandos que permitem o acesso ao conteúdo do banco de dados são incompreensíveis. Os robôs não estão programados para entender a estrutura de um banco de dados, ou as linguagens utilizadas para recuperar a informação.

No caso particular dos catálogos, como o *Yahoo*!, os *spiders* são capazes de armazenar as informações contidas em seus bancos de dados seguindo cada *link* das diversas categorias, num trabalho bastante árduo. Navegando através das hierarquias, o robô replica todos os conteúdos que resultariam das possíveis opções de busca do usuário.

Alguns exemplos de banco de dados podem ser vistos em *http://plants.usda.gov*, mantido pelo Departamento de Agricultura dos Estados Unidos, que oferece informações sobre plantas ou em http://www.tecepe.com.br/olimpiadas, com informações sobre as olimpíadas.

Os mecanismos de busca tradicionais podem encontrar somente a página inicial destes bancos de dados, mas não informações sobre o seu conteúdo. Ironicamente, estes recursos representam algumas das mais valiosas informações disponíveis na *Web*. Qualidade é a principal razão para explorar tais base de dados.

f) Políticas de Exclusão

Os *site*s que compõem o banco de dados dos buscadores tradicionais são, no geral, estáticos. Páginas estáticas são identificadas por um único endereço, URL, e são mostradas ao usuário quando este endereço é digitado no navegador. Os *spiders* são capazes de encontrar páginas estáticas desde que hajam *links* apontando para elas, a partir de páginas conhecidas. Entretanto,

os mecanismos de busca limitam o número de páginas que coletam utilizando alguns critérios. Estas páginas, que não farão parte do banco de dados, não são parte da *Web* oculta. Elas são visíveis e poderiam ser identificadas pelos robôs, mas os administradores decidem excluí-las para reduzir seus custos de operação. Por exemplo, certos tipos de linguagem de programação - tais como *Flash, Schokwave, Word, WordPerfect*, arquivos executáveis e comprimidos, páginas formatadas em *Portable Document Format* (PDF)[12] etc., podem ser excluídas porque, além de aumentarem o custo de operação das ferramentas de busca, tem menor procura.

Os mecanismos de busca são altamente competentes e otimizados para trabalhar com páginas em textos e, mais exatamente, em textos codificados em *HyperText Markup Language* HTML. Documentos em HTML obedecem um formato simples[13]. A simplicidade do formato facilita o trabalho dos mecanismos de busca para administrar, controlar, estocar e recuperar a informação. Os problemas se iniciam quando o conteúdo não obedece este modelo simples de página *Web*.

Os arquivos em formato PDF, por exemplo, preservam a aparência dos documentos impressos sendo assim, bastante utilizados para disponibilizar artigos, jornais, livros etc. Seu armazenamento em bancos de dados exige mais recursos computacionais porque um arquivo neste formato pode ser composto de centenas ou mesmo milhares de páginas. Tecnicamente é possível incluir o conteúdo destes arquivos nas bases de dados das ferramentas de busca. Entretanto, os administradores decidem não despender tempo e recursos nesta tarefa porque a maioria dos documentos neste formato são técnicos ou acadêmicos, utilizados por uma parcela comparativamente pequena de pessoas e irrelevante para a maioria dos usuários.

Existem na Internet catálogos e listas de endereços da *Web* oculta que podem mostrar o caminho para este conteúdo tão relevante. Entre eles:

- *Direct Search* (http://www.freepint.com/gary/direct.htm)
- *The Invisible Web Catalog* (http://www.invisibleWeb.com)
- http://www.internets.com com *links* para mais de mil bancos de dados de interesse acadêmico
- http://www.completeplanet.com que oferece acesso a cerca de 103 mil bancos de dados.

Pode-se, ainda, localizar a página de interface dos bancos de dados digitando-se nos buscadores tradicionais o assunto de interesse e a palavra *database* ("assunto" and *database*).

Considerações finais

Os mecanismos de busca são fundamentais para a recuperação das informações na *Web*. Entretanto, uma busca eficiente depende da consonância entre dois aspectos: habilidade do usuário no uso das ferramentas de busca e a capacidade do buscador em, a partir de um termo ou conceito, compreender as necessidades do usuário e recuperar as informações adequadas.

Do ponto de vista do usuário, o conhecimento das diferentes formas de operação e peculiaridades de cada ferramenta, e o correto planejamento e

[12] O Google é, atualmente, o único mecanismo de busca que indexa documentos em formato PDF.

[13] Cada página é composta de duas partes: o cabeçalho e o corpo do texto, que são claramente separados no código fonte do HTML. No cabeçalho constam o título, disposto de forma lógica no alto da página. O corpo contém o conteúdo propriamente dito.

operacionalização da estratégia de busca são fundamentais para a recuperação das informações.

De modo geral, os usuários podem adotar os seguintes princípios para recuperação da informação na *Web*:

· Utilizar mais de um buscador nas pesquisas. Os resultados em diferentes buscadores apresentam baixa redundância.

· Encontrar os buscadores mais adequados às suas necessidades. Especializar-se nas suas formas de funcionamento para extrair todo seu potencial.

· Obter vantagem das diferenças existentes entre os catálogos e os programas de busca.

· Utilizar os metapesquisadores para obter uma visão geral dos bancos de dados dos diferentes mecanismos de busca.

· Localizar, se possível, pesquisadores especializados em sua área de interesse

· Não esquecer da *Web* oculta

Do ponto de vista das ferramentas de busca, seus desenvolvedores devem procurar ir além da mera identificação, a partir de um termo ou conceito, dos conteúdos que serão apresentados ao usuário. Considerando que a maioria dos usuários da Internet não tem as habilidades básicas na manipulação de bases de dados, devem ser desenvolvidos ferramentas que indiquem como podem ser utilizadas de forma mais eficiente.

Hawkins (1978), por exemplo, sugere que, a partir de um termo ou conceito oferecido pelo usuário, o banco seja capaz de identificar citações relacionadas e, a partir delas, extrair outros termos ou conceitos que indiquem novas estratégias de busca. Simon e Valdez-Perez (1997) estudam os programas interativos de busca que, a partir do título dos documentos recuperados, estabelecem entradas para outros documentos similares na base de dados.

Desenvolvimentos neste sentido são, sem dúvida, importantes. Entretanto, sua ação é limitada por estarem baseados na análise textual e não nos modelos humanos de busca. A maioria dos sistemas especialistas desenvolvidos para auxiliar os usuários finais na consulta às bases de dados são incapazes de processar outros critérios relativos ao pedido de busca.

Ferramentas mais eficientes poderiam ser construídas a partir da observação do comportamento dos usuários. Os buscadores deram um passo nesta direção quando notaram que, a maioria dos usuários, digitava dois ou mais termos nos campos de busca sem utilizar nenhum comando entre as palavras. A expectativa dos usuários era de recuperar documentos com todos os termos digitados. Mas, os buscadores apresentavam, além destas páginas, outras onde constavam apenas um dos termos de busca.

Por exemplo, digitando-se a expressão *comércio exterior* no campo de busca, o usuário recuperava, além dos documentos com a expressão, outros onde ocorriam apenas a palavra *comércio e* a palavra *exterior.* Este procedimento gerava um grande número de resultados sem relevância. Os desenvolvedores passaram, então, a incluir automaticamente o comando + entre os termos, recuperando somente os documentos com todos os termos de busca.

A observação do comportamento e a avaliação das expectativas dos usuários permitiu que os desenvolvedores – a partir de alterações técnicas simples – tornassem suas ferramentas significativamente mais eficientes.

A maioria dos usuários dos bancos de dados na Internet é inexperiente, não tem conhecimento dos controles básicos e não explora adequadamente todo o potencial dos buscadores. Analisando o comportamento de busca, ou seja, o que as pessoas fazem e pensam quando estão buscando informações na Internet, os desenvolvedores poderiam mapear e automatizar suas rotinas. A criação de sistemas que induzissem o usuário ao comportamento adequado e/ou que adequassem suas necessidades às peculiaridades da ferramenta aumentaria, de forma significativa, a eficiência na recuperação das informações. Assim, o caminho para tornar estas ferramentas cada dia mais amigáveis está, principalmente, na observação do comportamento do usuário leigo que freqüenta a Internet.

Finding information on the Web

There are thousands of pages on a great variety of subjects and interests on the Web. Finding information, however, is not a trivial task. Search tools are important in assisting us in this task. This text intends to show the differences of operation among several currently existing searchers on the Internet and how their peculiarities affect the results. Knowing their characteristics and they way they function it is possible to take advantage of all the potential of each tool and, thus to find the desired information more efficiently. We will also comment on the so-called invisible Web, a great amount of information that is not made available by traditional search tools.

Key-words: Internet; *Web;* Search tools; Search Engines; Directories; Meta-search Engines; Invisible *Web*

Referências

BARKER J. (Coord.) *Find information on the Internet: a tutorial*. Disponível em WWW. URL: *http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfor.html*, acesso em 7 ago. 2002.

BERGMAN, M. K. *The deepWeb surfacing hidden value.* Disponível em WWW. URL: *http://www.press.umich.edu/jep/07-01/bergman.html*, acesso em 13 jun. 2002.

HAWKINS, D.T. Multiple database searching: techniques and pitfalls. *Online*, v. 2, n. 2, p. 1-15, 1978.

LAWRENCE S.; GILES C.L. Accessibility of information on the Web. *Nature*, London, 400, 107, 1999.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. *Ciência da Informação*, Brasilia , v. 31, n. 2, p. 60-71, maio/ago. 2002.

NOTESS G.R. *Searching the hidden Internet*. Disponível em WWW. URL: *http://www.onlineinc.com/database/JunDB)//ntes6.html*, acesso em 5 jun.1997.

_____, Internet search techniques and strategies. Disponível em WWW. URL: *http://www.onlineinc.com/onlinemag/Jul0197/net7.html*, acesso em 5 jun.1997.

OLDROYD, B K; CITROEN, C L Study of strategies used in online searching. *Online Review,* v. 1, n. 4, p. 295-310, 1997.

SIMON, H. A.; VALDEZ-PEREZ, R. E. Scientific discovery and simplicity of method. *Artificial Intelligence*, v. 91, n. 2, p. 183-203, Apr. 1997.

SHERMAN C. *The invivsible Web.* Disponível em WWW. URL: *http://www.freepint.co.uk/issues/0806000.htm*, acesso em 8 ago.2001

SULLIVAN, D. (Ed.). *Search engine features for searchers*. Disponível em WWW. URL: *http://www.searchenginewatch.com/facts/ataglance.html*, acesso em 3 jun. 2002.

_____, *Power searching for anyone*. Disponível em WWW. URL: *http://www.searchenginewatch.com/facts/powersearch.html*, acesso em 26 out. 2001.

_____, Search engine math. Disponível em WWW. URL: *http://www.searchenginewatch.com/facts/math.html*, capturado em 26 out. 2001.

WISEMAN K., *The invivisble Web.* Disponível em WWW. URL: *http://www3.dist214.k12.il.us/invisible/article/invisiblearticle.html*, acesso em 6 maio 2000.