Ferramentas de busca na *Web*

Beatriz Valadares Cendón
Professora adjunta da Escola de Ciência da Informação da Universidade Federal de Minas Gerais
Cendon@eb.ufmg.br

Resumo

Existem hoje centenas de ferramentas para busca de informações nas cerca de um bilhão de páginas HTML que se estimam existir na Web. As peculiaridades destas ferramentas influenciam no tipo, número e qualidade dos recursos recuperados através delas. Este artigo oferece uma visão das principais categorias de ferramentas de busca da Internet, suas semelhanças, diferenças e características, bem como analisa as vantagens e desvantagens associadas a cada uma, de forma a proporcionar ao profissional da informação instrumental para aumentar sua eficiência na procura de recursos informacionais.

Palavras-chave

Internet, Ferramentas de busca; Web; Diretórios; Motores de busca; Metamotores

Web search tools

Abstract

Nowadays there are hundreds of different tools for searching the estimated one billion Web pages. Their peculiarities influence the type, number and quality of resources retrieved through their use. This article offers an overview of the main categories of web search tools, their similarities, differences and characteristics in order to provide the information professionals an instrument to improve their efficiency in the search for information.

Keywords

Internet, Search tools; Web; Directories; Motors of search; Metamotors

Desde os primórdios da Internet, houve a preocupação de se criarem ferramentas para localização de seus recursos informacionais. Entre as ferramentas mais antigas, podem-se citar o *Archie*, que busca arquivos em repositórios de FTP, e *Veronica* e *Jughead*, que encontram conteúdos armazenados nos *Gophers*. Com o advento da *Web* e a conseqüente explosão das publicações disponibilizadas por meio dela, começaram a surgir as ferramentas específicas para pesquisa de suas páginas. Existem hoje centenas destes instrumentos que fornecem meios para localizar o que se busca entre as cerca de um bilhão de páginas HTML, que se estimam.

Existem dois tipos básicos de ferramentas de busca na *Web*: os motores de busca e os diretórios. Entretanto, a partir dessas duas categorias básicas, outros tipos de ferramentas têm surgido, fazendo o mundo dos serviços de busca complexo e volátil. Devido às características específicas de cada ferramenta, o tipo, número e a qualidade dos recursos recuperados através de seu uso, podem variar enormemente. Para obter melhores resultados na busca de informações, o primeiro passo é entender as peculiaridades dos diferentes tipos de ferramentas de busca na *Web*. Este artigo oferece uma visão das principais categorias de ferramentas de busca da Internet, suas semelhanças, diferenças e características e analisa, também, as vantagens e desvantagens associadas a cada uma, de forma a proporcionar ao profissional da informação instrumental para aumentar sua eficiência na procura de recursos informacionais.

DIRETÓRIOS

Os diretórios foram a primeira solução proposta para organizar e localizar os recursos da *Web*, tendo precedido os motores de busca por palavras-chave. Foram introduzidos quando o conteúdo da *Web* ainda era pequeno o suficiente para permitir que fosse coletado de forma não automática. Organizam os *sites* que compõem sua base de dados em categorias, as quais podem conter subcategorias, ou seja, os *sites* recebem uma organização hierárquica de assunto e permitem aos usuários localizar informações, navegando, progressivamente, para as subcategorias. Como são ferramentas genéricas, destinadas a um público variado, procuram incluir, em suas árvores hierárquicas de assunto, tópicos que são de interesse amplo. É comum que incluam, por exemplo, itens relacionados com educação, esporte, entretenimento, viagens, compras ou

informática. Cabeçalhos de assunto são atribuídos de forma consistente, de modo que os usuários podem contar com a ajuda de um vocabulário controlado.

Os *sites* coletados passam pela seleção, na maioria das vezes, por seres humanos, os editores, que tomam conhecimento de novos recursos por meio de sugestões de usuários, de pesquisas na Internet (em listas de anúncios de novas páginas e atualizações, por exemplo), ou ainda, pelo uso de robôs para coletar novos URLs. O número de editores empregados, que pode variar de 30 (utilizados pelo Snap*) a mais de 15 mil (como no caso do Open Directory da Netscape), é um sinal da qualidade e atualização dos dados, mas não uma garantia. Embora normalmente os critérios para seleção utilizados não sejam divulgados, apenas os melhores recursos são escolhidos para inclusão. Apesar desta triagem, devido à enorme quantidade de sugestões, centenas de *sites* podem ser acrescentados semanalmente. Os grandes diretórios podem conter dezenas de milhares de categorias e subcategorias e mais de um milhão de *sites*.

O primeiro diretório da *Web* foi o The World Wide Web Virtual Library (http://www.vlib.org/), lançado em novembro de 1992 e sediado no CERN, que também foi o local de nascimento da *Web*. Atualmente, o exemplo mais conhecido é o Yahoo!, que iniciou em 1994, a partir de um *hobby* de estudantes de doutorado na Stanford University, e hoje é uma bem-sucedida empresa comercial. Outros exemplos de diretórios são Snap (http://www.snap.com), LookSmart (http://www.looksmart.com), Open Directory (http://dmoz.org/), Yahoo Brazil (http://www.br.yahoo.com), Cadê (http://www.cade.com.br), Surf (http://www.surf.com.br) e Vai & Vem (http://www.vaievem.com.br), sendo estes três últimos brasileiros.

DIFERENÇAS ENTRE OS DIRETÓRIOS

Embora todos os diretórios sigam os princípios genéricos descritos anteriormente, variam quanto aos princípios de organização, à forma de descrição dos recursos e aos assuntos cobertos, apresentando características próprias.

Quanto aos princípios de organização, a maioria dos diretórios usa as listas hierárquicas de assunto. Entretanto, alguns utilizam esquemas tradicionais de classificação, como o sistema de cabeçalhos de assunto da Library of Congress, utilizados pelo diretório do Scout Reports (http://www.signpost.org/signpost/), ou a classificação Dewey, usada pelo BUBL Link (http://bubl.ac.uk/link/). Geralmente, estes são criados e mantidos por profissionais da informação ou bibliotecários, em uma tentativa de promover melhores formas de acesso aos recursos da *Web*.

Quanto às descrições dos *sites*, a maior parte dos diretórios que se constituem em empresas comerciais limita-se a incluir títulos e breves resumos de até 30 palavras. Alguns diretórios se diferenciam dos demais por fornecer descrições criteriosas e detalhadas dos recursos, podendo incluir críticas e avaliações dos mesmos. Para elaboração das análises, estes **diretórios avaliativos ou acadêmicos** utilizam estudantes de mestrado ou mestres em biblioteconomia e ciência da informação, ou ainda especialistas em assuntos específicos. São geralmente associados a bibliotecas ou instituições de ensino, utilizam um processo seletivo de recursos mais rigoroso e não incluem propaganda. Porém, são poucos os diretórios que se enquadram nesta categoria. Dentre eles, destaca-se, por sua qualidade, o Argus (http://www.clearinghouse.net/), que iniciou como um projeto da University of Michigan e é agora gerenciado por profissionais da informação. Coleta apenas *sites* que são guias de recursos na *Web* sobre um determinado assunto, os quais são compilados por especialistas em seus campos e fornecem *links* relevantes na área coberta. Cada guia é avaliado pela equipe do Argus, que os classifica em uma de suas 13 categorias principais e lhes atribui nota de 1 a 5, de acordo com vários critérios de qualidade, como *design*, conteúdo e outros. O Argus apresenta uma detalhada descrição de suas políticas de seleção e classificação dos *sites*. Outros exemplos de diretórios avaliativos são o Infomine (http://infomine.ucr.edu), o Britannica.com (http://www.britannica.com) e o Scout Reports Signpost e a WWW Virtual Library, mencionados anteriormente.

Quanto aos assuntos, nem todos os diretórios são genéricos como o Yahoo!, ou o Britannica, que cobrem todos os assuntos. Alguns diretórios cobrem áreas específicas e têm sido chamados de **diretórios temáticos ou especializados**. Existem, por exemplo, diretórios especializados em imagens, jornais e revistas, *software*, listas de discussão; outros coletam *sites* sobre assuntos específicos como saúde, ciências, legislação, informática etc.; ainda outros listam ferramentas de busca de países específicos ou para um público-alvo determinado (crianças, pesquisadores, organizações não-governamentais etc.). Alguns *sites* se especializam em listar estes diretórios temáticos, como, por exemplo, Tematicos (http://www.tematicos.com), Buscopio (http://www.buscopio.com), Beaucoup (http://www.beaucoup.com) ou o Search Engine Watch (http://www.searchenginewatch.com/links/Specialty–Search–Engines/). Para ferramentas

* Em setembro de 2000, o Snap mudou o nome para NBCi (http://www.nbci.com/)

regionais, veja-se também o *site* do Search Engine Watch (http://www.searchenginewatch.com/links/Regional–Search–Engines/).

Devido à freqüência com que novas ferramentas de busca (diretórios e motores de busca) surgem, ao mesmo tempo em que outras caem em desuso, um novo tipo diretório passou a ser criado: os **diretórios de ferramentas de busca**. Diretórios como FinderSeeker (http://www.finderseeker.com/) ou Search.com (http://search.cnet.com/) têm o objetivo de listar ferramentas de busca, para facilitar sua identificação.

MOTORES DE BUSCA

Ao contrário dos diretórios, os motores de busca não organizam hierarquicamente as páginas que colecionam. Preocupam-se menos com a seletividade que com a abrangência de suas bases de dados, procurando colecionar o maior número possível de recursos através do uso de *softwares* chamados robôs. Como suas bases de dados são extremamente grandes, podendo alcançar centenas de milhões de itens, permitem aos usuários localizar os itens desejados mediante buscas por palavras-chave, ou, às vezes, em linguagem natural.

Os motores de busca começaram a surgir quando o número de recursos na *Web* adquiriu proporções tais que impediam a sua coleta por meios manuais e a busca apenas através da navegação. A maioria deles derivou do trabalho de estudantes de pós-graduação, professores, funcionários do departamento de sistemas de empresas ou outras pessoas interessadas na *Web*. Muitos não obtiveram continuidade, à medida que a tarefa a ser executada passou a exigir maiores recursos humanos e técnicos. Os que sobreviveram foram adquiridos por empresas ou financiados por propagandas, investidores e recursos de pesquisa.

ALIWEB (Archie-Like Indexing on the Web) e Harvest são exemplos das primeiras tentativas de criar motores de busca por palavras-chave, e utilizavam tecnologias diferentes das atuais. O primeiro dos motores baseados em robôs foi o WebCrawler, lançado em abril de 1994. Todos os motores atuais utilizam o método de robôs sendo formados por quatro componentes: um robô, que localiza e busca documentos na *Web*; um indexador, que extrai a informação dos documentos e constrói a base de dados; o motor de busca propriamente dito; a interface, que é utilizada pelos usuários.

Os **robôs**, também chamados de aranhas (*spiders*), agentes, viajantes (*wanderers*), rastejadores (*rawlers)* ou vermes (*worms)*, são programas que o computador hospedeiro da ferramenta de busca lança regularmente na Internet, na tentativa de obter dados sobre o maior número possível de documentos para integrá-los, posteriormente, à sua base de dados. Existem várias estratégias que os robôs podem utilizar para se locomoverem de um documento a outro, utilizando-se dos *links* existentes nas páginas da *Web*. Geralmente, eles iniciam a busca a partir de *sites* conhecidos, especialmente daqueles que possuem muitos *links*, recuperam a sua *home page* e, sistematicamente, seguem os *links* encontrados nesta página inicial. Usam algoritmos próprios para determinar que *links* devem seguir. Por exemplo, alguns recuperam os documentos da hierarquia superior de um grande número de servidores (abordagem *breadth-first*), enquanto outros capturam todos os documentos em *links* de um mesmo servidor (abordagem *depth-first*).

Os motores de busca podem usar vários robôs que trabalham em paralelo para construir sua base de dados. Por exemplo, o Excite empregava, no começo do ano 2000, cerca de 10 aranhas para pesquisa na rede. Ela anunciou que deverá acrescentar outra dezena delas, cada uma com a capacidade para cobrir 50 milhões de páginas da Internet. Na coleta de páginas para suas bases de dados, a maioria dos motores de busca permite também que os usuários sugiram URLs, em vez de esperar que os documentos sejam encontrados através da varredura realizada regularmente pelos robôs.

Os documentos encontrados pelos robôs são encaminhados aos **indexadores** que extraem a informação das páginas HTML e as armazenam em uma base de dados. Esta base de dados do motor de busca consiste de informações julgadas importantes como os URLs ou endereços das páginas HTML, títulos, resumos, tamanho e as palavras contidas nos documentos.

A **interface**, normalmente uma página *Web*, é utilizada pelos usuários para efetuar a pesquisa na base de dados. Fornece meios para que o usuário formule a sua consulta, que é recebida e transmitida para o *software* de busca ou **motor de busca** propriamente dito. Este é um programa que localiza, entre os milhões de itens na base de dados, aqueles que devem constituir a resposta. O programa também é responsável pela ordenação dos resultados, de maneira que os mais relevantes apareçam em primeiro lugar na lista de resultados. Os resultados mostrados contêm uma lista de descrições de *sites* e seus *links*.

DIFERENÇAS ENTRE OS MOTORES DE BUSCA

Todos os motores de busca são compostos dos componentes listados anteriormente. Entretanto, diferem entre si em relação a fatores como o tamanho de suas bases de dados, critérios para indexação e inclusão de páginas, além de ordenação dos resultados. Suas interfaces, recursos de busca que oferecem, a freqüência com que atualizam suas bases de dados e o modo como apresentam os resultados também variam.

Embora aqui o foco principal seja nos motores genéricos, é importante observar que, da mesma forma como existem diretórios temáticos, existem também **motores de busca temáticos**, que se especializam em um determinado tópico. Veja-se, por exemplo, o Medical World Search (http://www.mwsearch.com), que se especializa em encontrar informações médicas. Os diretórios de ferramentas de busca, já citados, permitem localizar estes motores temáticos.

Tamanho da base de dados

O tamanho das bases de dados dos motores de busca é medido, geralmente, em número de URLs. Este tamanho é de alta relevância para que a ferramenta seja considerada boa, já que os recursos informacionais na Internet só podem ser encontrados em uma pesquisa, se alguma ferramenta os tiver incluído. Se um motor cobre mais da *Web*, ele terá maior chance de conter a informação procurada. Conseqüentemente, os motores maiores tendem a ser mais usados, atraindo maior número de anunciantes e podendo cobrar maiores taxas pelos anúncios.

Entretanto, nenhum motor de busca contém todas as páginas existentes na *Web*. Os melhores não chegam a incluir 60% delas, como mostra a tabela 1. Nela estão listados os maiores motores do mundo, com o número de páginas em suas bases de dados e a percentagem do número total de páginas da *Web* que cada um indexa.

Entre os motores estrangeiros, o Altavista e HotBot (que usa na realidade uma base de dados compilada pelo serviço Inktomi) destacaram-se por vários anos como sendo os maiores do mundo. Mais recentemente, quatro motores, WebTop.com, Fast Search, Google e Northern Light, têm despontado na competição. Dentre os motores que indexam unicamente *sites* brasileiros, destaca-se o Todobr. Lançado em novembro de 1999 e com tecnologia desenvolvida pela Universidade Federal de Minas Gerais, ele continha, em junho de 2000, cerca de 10 milhões de páginas, ou seja, quase a totalidade da *Web* brasileira. Para páginas do Brasil, costuma trazer mais resultados que as maiores ferramentas estrangeiras.

TABELA 1
Tamanho da base de dados dos motores de busca

Motor de busca	Nº de páginas (em milhões)	% da Web
Google	560	56%
WebTop.com	500	50%
Altavista	350	35%
Fast	340	34%
Northern Light	265	27%
Excite	250	25%
HotBot / Inktomi	110	11%
Go / Infoseek	50	5%
Lycos	50	5%

Fonte: Search Engine Watch. *Search engine sizes*. Disponível na Internet via WWW. URL: http://searchenginewatch.com/reports/sizes.html. Arquivo capturado em 29/set./2000.

Embora gigantescas, as bases de dados de cada motor não são iguais. Assim, para a mesma busca, cada mecanismo invariavelmente trará bons resultados que outros não encontraram. Para uma busca ser completa, necessariamente há de se usar mais de uma ferramenta.

Critérios para indexação

Os motores de busca criam índices, chamados, na linguagem técnica, de arquivos invertidos, que são utilizados para dinamizar a busca de informações na sua base de dados. No índice, são inseridos todos os termos que podem ser utilizados em busca de informações e o URL das páginas que os contêm. A fim de fornecer melhores recursos para recuperação dos resultados e sua ordenação, podem ser ainda armazenados dados sobre a posição das palavras na página e sobre os *tags* HTML associados com o texto. Se um termo não estiver incluído no índice, ele não será encontrado, portanto os critérios utilizados para indexação influenciam os resultados das buscas.

A maioria dos motores de busca indexa, ou seja, inclui, em seu índice, cada palavra do texto visível das páginas. Entretanto, alguns extraem, em vez do texto completo, apenas o URL, as palavras que ocorrem com freqüência, ou palavras e frases mais importantes contidas no título ou nos cabeçalhos e nas primeiras linhas, por exemplo. Alguns motores indexam também outros termos, que não fazem parte do texto visível, mas que contêm informações importantes e úteis. Exemplos deste tipo de texto são os textos incluídos nos *metatags* para classificação, descrição e palavras-chave e texto ALT do *tag Image*, ou seja, texto associado com imagens. Os *metatags* de classificação

fornecem uma palavra-chave que define o conteúdo da página. Os de descrição retornam à descrição da página feita pelo seu autor no lugar do resumo que o robô criaria automaticamente. Os de palavras-chave fornecem as palavras-chave designadas pelo autor para descrever seu conteúdo ou assunto. Por exemplo, no *metatag* <*META name=”keyword” content=“Brasil, informação para negócios”* >., as palavras Brasil e informação para negócios podem não fazer parte do texto visível da página, entretanto foram indicadas pelo seu autor como indicadores do assunto sobre os quais a página versa.

Alguns motores não incluem no seu índice algumas palavras do texto, chamadas palavras proibidas (*stop words*). Palavras proibidas são selecionadas entre as muito comuns, como, por exemplo, a preposição “de”, ou o artigo “the” na língua inglesa. Como ocorrem nos textos em alta freqüência, muitos motores as excluem em seus índices para economizar espaço de armazenamento. Outros as incluem nos índices, mas os ignoram ao fazer uma busca, para torná-la mais rápida. Para o usuário, isto é problemático, uma vez que os motores, em geral, não fornecem documentação sobre quais são as palavras proibidas utilizadas.

Critérios para inclusão de páginas

O número de itens nas bases de dados dos motores é determinado pelos critérios que utilizam para inclusão de páginas. Alguns motores procuram incluir todas ou a maioria das páginas de cada *site* visitado. Outros indexam os *sites* superficialmente, ou seja, incluem apenas a *home page* e algumas páginas principais. Além de documentos HTML, são cada vez mais comuns motores que coletam e indexam outros formatos, como imagens, vídeos, gráficos, arquivos PDF ou ASCII. Outros compilam ainda mensagens em grupos de discussão, *sites* de FTP, *menus* de *gophers* e outros recursos.

Entretanto, existem páginas que não são parte de nenhum motor de busca. Estas incluem *sites* que requerem senhas para entrada, páginas atrás de uma *firewall* e páginas que contenham o *metatag* Meta Robot “noindex”. O *metatag* Robot (<META name=“robots” content=“noindex”) pode ser acrescentado aos marcadores de cabeçalho pelo criador da página para indicar aos robôs que eles não devem capturá-la. Páginas isoladas, que não sejam referenciadas através de *links* em outras páginas na Internet, também podem escapar à varredura dos robôs.

Existe ainda uma parte da *Web* que tem sido chamada de *Web* invisível, por incluir páginas não indexadas pela maioria dos motores de busca. Parte da *Web* invisível são as páginas que contêm *frames*, *image-maps* e as páginas dinâmicas. No caso de páginas que contenham *frames*, é comum ver *sites* com mais de 100 páginas terem apenas sua *homepage* indexada. Altavista, Google, Fast e Northern Light são alguns dos poucos motores que indexam *frames*, mas, mesmo assim, não o fazem de maneira ideal, pois não trazem o contexto em que elas estão inseridas. Páginas que usam *frames* muitas vezes são planejadas de forma que, para o seu entendimento, é necessário visualizar o conjunto das informações. Ao mostrar uma *frame* fora de seu contexto, os *links* para navegação para o restante do *site* podem não estar presentes, aprisionando o usuário àquela página, ou simplesmente as páginas podem não fazer sentido.

Páginas dinâmicas também representam um desafio para os robôs. Geralmente elas são formadas de informações contidas em bases de dados, e são montadas no momento em que o usuário clica em um *link*, ou seja, as páginas são criadas no ato da busca. Caracterizam-se por conter, geralmente, um ponto de interrogação como parte do seu URL. Por exemplo: um URL de uma página dinâmica poderia ser algo do tipo: http://www.website.com/cgi-bin/getpage.cgi?name=sitemap. A maioria dos motores de busca, ao encontrar o ponto de interrogação no endereço, recusam a indexação destas páginas, para evitar situações em que eles obteriam milhares de páginas, quase iguais, porém com URLs ligeiramente diferentes. Isso se torna um problema, na medida em que páginas dinâmicas têm sido crescentemente utilizadas na Internet. Algumas ferramentas que se especializam em dar acesso a informações contidas em bases de dados não indexadas por nenhum motor de busca são InvisibleWeb.com (http://www.invisibleweb.com/), Lycos Invisible Web Catalog (http://dir.lycos.com/Reference/Searchable_Databases/ e Direct Search (http://gwis2.circ.gwu.edu/~gprice/direct.htm).

Da mesma forma, os motores podem não indexar páginas relacionadas a *image-maps*. *Image-maps*, também chamados de mapas de imagem ou mapas clicáveis, consistem de uma figura contendo dois ou mais *links*, cada um vinculado a uma região da imagem. Alguns dos motores que o fazem são AltaVista, Go, e Northern Light.

Freqüência de atualização dos dados

Devido ao dinamismo da Internet, as bases de dados dos motores de busca precisam ser atualizadas, não só para adicionar novas páginas, mas também para deletá-las ou incluir as modificações das já existentes no índice. Caso os robôs não revisitem periodicamente toda a Internet, os URLs que eles trazem como resultados de uma busca podem não mais existir, ou podem existir, mas não mais conter as mesmas informações, e, portanto, não mais serem relevantes para a busca.

Os motores de busca se propõem a atualizar completamente seus índices pelo menos uma vez por mês. Partes mais importantes desses, como, por exemplo, páginas mais populares entre os usuários (Excite, Lycos), ou páginas que mudam com mais freqüência (Inktomi, Infoseek, Altavista, Go), podem ser atualizadas assiduamente, em torno de uma vez por semana, enquanto o restante do índice é atualizado a cada duas a quatro semanas. Novos URLs e *links* mortos descobertos pelos robôs são atualizados diariamente. Cada motor tem sua própria estratégia e tecnologia para se manter atualizado, embora possa acontecer que algum deles passe alguns meses sem acrescentar novos URLs ou modificar seus índices.

Os motores diferem também quanto ao tempo necessário para que uma página coletada pelos robôs ou submetida pelos usuários seja adicionada ao índice. Até que isso aconteça, a informação não será encontrada através de pesquisa no motor. Com o crescente número de *sites* disponíveis na Internet e a concorrência para chamar a atenção das ferramentas de busca, podem se passar meses antes que um *site* novo seja adicionado à base de dados. Empresas especializadas em buscas na Internet, como a LookSmart e a Inktomi, estão começando a disponibilizar programas que cobram uma taxa dos *sites* de Internet para disponibilizá-los em um prazo de 48 horas após a solicitação.

Interfaces e recursos para busca

Os motores diferem também em relação às interfaces e recursos de busca que oferecem. Geralmente fornecem dois modos de busca, a busca simples para usuários leigos e a busca avançada para usuários mais experientes ou profissionais. Na busca simples, existem janelas e *menus* que permitem que os usuários entrem nos termos de busca sem a necessidade de conhecimento de lógica booleana. A busca avançada fornece recursos mais poderosos, como expressões booleanas complexas. Muitas vezes, na busca simples, os conectivos booleanos são automaticamente colocados entre os termos de busca, e nem sempre os usuários sabem qual operador está sendo utilizado. Em alguns motores, por exemplo, um espaço entre os termos da consulta é interpretado como um conectivo booleano OR (Altavista e Excite, por exemplo), enquanto para outros tem o significado de AND (Google e Northernlight, por exemplo).

Podem oferecer recursos como truncamento, busca por frase, busca por proximidade de palavras, busca por campos e sensibilidade à caixa de caracteres (isto é, caixa-alta e caixa-baixa). É comum também haver opções para permitir a limitação por data, domínio, idioma ou tipo de arquivos (com base na extensão dos nomes dos arquivos). Alguns motores fornecem opções mais sofisticadas, como a busca automática pela raiz das palavras, ou seja, se o usuário entrar com a palavra "psicologia", ele encontrará também documentos com a palavra "psicólogo". Em alguns casos, a pesquisa se estende também a outros termos sinônimos ou a termos com conteúdo semântico equivalente ao termo da consulta, como é o caso do Excite. Esta busca estendida, quando existente, é geralmente automática, não sendo dada ao usuário a possibilidade de desabilitá-la. São mais raros motores que permitem buscas em linguagem natural, na qual a consulta pode ser entrada na forma de uma sentença, em vez de termos isolados.

Não existe ainda uma completa padronização nas interfaces e recursos de busca que cada mecanismo oferece, os quais variam de motor para motor. Para se usar corretamente cada motor, é necessária a leitura das páginas de ajuda ou a consulta a tabelas comparativas em revistas especializadas ou na própria Internet (ver, por exemplo, o *site* da biblioteca da University of California at Berkeley – http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/ToolsTables.html e o *site* da University at Albany Library - http://www.albany.edu/library/internet/choose.html).

Critérios de ordenação dos resultados

Devido à quantidade de páginas na Internet, na maioria das vezes obtém-se um grande número de resultados para qualquer busca. Portanto, a seqüência em que os resultados são mostrados torna-se importante. Se duas ferramentas trazem o mesmo número de resultados, porém uma delas traz itens mais relevantes entre os primeiros resultados, ela será considerada melhor. Com a finalidade de permitir que os melhores *sites* apareçam em primeiro lugar, a maioria dos motores de busca utiliza algoritmos de ordenação de resultados.

Entre os critérios mais utilizados por estes algoritmos estão a localização e freqüência de ocorrência das palavras em uma página. Por exemplo, se o termo de busca aparece no título, em cabeçalhos de destaque ou nos primeiros parágrafos em uma página, esta seria considerada mais relevante que outras páginas em que as palavras de pesquisa não aparecem nestas posições. E, se uma palavra aparece com mais freqüência em uma página que em outra, a primeira seria considerada mais relevante.

Outros critérios comuns para a determinação de relevância incluem o número de termos da consulta que estão presentes na página e a proximidade em que os termos se encontram. Os motores podem também levar em consideração o tamanho do documento. Se dois documentos possuem o mesmo número de ocorrência dos termos de busca, os documentos mais curtos seriam considerados mais importantes que documentos longos. Este critério é chamado de densidade, pois mede a densidade com que um dado termo é usado em cada documento. Às vezes aplica-se uma curva declinante, em que a primeira ocorrência de um termo conta mais que a segunda, que conta mais que a terceira, e assim por diante.

Os motores podem levar também em consideração o número total de vezes que uma palavra ocorre na base de dados, pois existe uma relação inversa entre o conteúdo informacional de um termo e o número de vezes que este ocorre em um texto. Assim, as palavras recebem pesos inversos à freqüência de sua ocorrência na totalidade dos documentos indexados pelo motor. Ou seja, palavras de ocorrência muito comum podem receber um peso menor que palavras relativamente raras na base de dados. O mecanismo exato para determinação da importância das páginas varia de motor para motor e geralmente não é revelado, porque os algoritmos de ordenação por relevância são um dos maiores fatores diferenciais de competição entre os motores.

Alguns motores de busca permitem que o usuário altere a ordenação dos resultados com critérios pessoais especificando, por exemplo, a necessidade da presença de todos os termos da busca, ou outros termos que possam determinar uma classificação mais alta do documento.

Por ser um fator tão importante, os motores têm desenvolvido novos mecanismos de ordenação, além daqueles estatísticos mencionados anteriormente. Alguns destes métodos que têm sido utilizados mais recentemente são:

- *Metatags* de palavra-chave e descrição

Alguns motores atribuem maior relevância às páginas que contêm os termos de busca nos *metatags* de descrição ou nos *metatags* de palavra-chave. Este método pressupõe que os autores utilizarão criteriosamente estes *tags*, o que pode não ocorrer. Alguns autores podem utilizar estes *metatags* para chamar atenção sobre suas páginas, utilizando termos que não correspondem ao seu conteúdo.

- Popularidade dos *links*

Popularidade dos *links* refere-se ao número de *links* que apontam para uma página. Alguns motores, consultando sua própria ou outras bases de dados, atribuem maior relevância a páginas muito referenciadas em outros *sites* ou referenciadas em *sites* importantes. *Links* são vistos como um voto sobre a qualidade das páginas. Um motor que tem destacado este aspecto nos seus critérios de ordenação por relevância é o Google.

- Direct Hit

Direct Hit (http://www.directhit.com) é um serviço na *Web* que monitora quais os *links* que milhares de usuários selecionam entre os resultados apresentados para uma busca e quanto tempo permanecem nos *sites* selecionados. Na ordenação por relevância, Direct Hit leva em consideração aqueles itens preferidos e selecionados por um grande número de internautas para a mesma consulta, ou consultas parecidas. HotBot, Lycos (http://www.lycos.com) e Metabusca (http://www.metabusca.com), por exemplo, são motores de busca que usam os serviços de DirectHit para trazer outras alternativas, além das normalmente encontradas na busca em sua base de dados, e, sua função é atribuir relevância mais alta às páginas mais visitadas pelos usuários.

- Inclusão do *site* em diretórios

Os motores híbridos, ou seja, aqueles que possuem um diretório vinculado ao *site*, costumam atribuir maior relevância aos *sites* selecionados para inclusão no seu diretório, por existir uma probabilidade de que estes sejam mais importantes para a consulta. Motores que têm adotado esta estratégia incluem Altavista, Infoseek e Lycos.

• Conceitos

O motor Northern Light aplica uma análise conceitual aos termos da consulta para determinar a intenção da busca. Esta análise é feita com o uso de índices gerados por seres humanos. Organiza os resultados em pastas que representam conceitos ou assuntos, tipos de *sites* (*press-releases*, mapas), ou idiomas. Os resultados dentro de cada pasta podem ser agrupados em novas pastas e finalmente são ordenados por relevância. Este sistema permite ao usuário ignorar as pastas irrelevantes e escolher apenas as que melhor se adeqüem à pergunta.

• Pagamento

Um motor, o Go To (http://www.goto.com), diferencia-se dos outros por apresentar em primeiro lugar os *sites* cujos produtores pagaram para estar entre eles.

• *Spam*

Spam pode ser definido como um conjunto de métodos considerados pouco éticos para promover páginas através da repetição da palavras irrelevantes, mas muito procuradas (como, por exemplo, futebol), para que estas páginas, embora não relacionadas com a consulta, sejam localizadas por buscas comuns. Técnicas usuais de *spam* são o uso de texto invisível (texto escrito da mesma cor do fundo da página e que, portanto, apesar de poder ser lido pelos robôs, não é visto pelo usuário), texto escrito em letras muito pequenas, que também são difíceis de ser vistas, ou a inclusão de palavras não apropriadas nos *metatags*. Alguns robôs podem detectar esta repetição desnecessária de palavras e penalizar a página na ordenação por relevância, ou mesmo excluí-las do seu índice.

• *Forma de apresentação dos resultados*

Os motores podem fornecer várias opções de formato de exibição à escolha do usuários. A maioria apresenta o número total de documentos encontrados, os quais são exibidos em páginas sucessivas contendo em torno de 10 resultados por página em um formato *default*, o qual mostra o título e um pequeno resumo. Normalmente o número de resultados por página pode ser alterado pelo usuário. Outras vezes, pode-se também determinar quantos documentos, do total encontrado, deseja-se receber. Por exemplo, podem ter sido encontrados 200 documentos, mas o usuário pode solicitar a apresentação apenas dos 10 ou 20 primeiros. O formato de exibição pode incluir o título, resumo, tamanho do arquivo em *bytes*, data do arquivo, URL e idioma. Alguns motores agrupam os resultados por URL e outros oferecem opção de apresentação de resultados de forma resumida ou detalhada. Raramente se oferece o recurso de destaque (*highlighting*) dos termos da consulta nos resultados apresentados.

Outros recursos de busca podem estar presentes na apresentação de resultados, como, por exemplo, *more like this* (usado pelo Excite), ou *related pages* (usado pelo Altavista), para permitir aos usuários a identificação de outros documentos semelhantes ao original. Ainda outras ferramentas podem apresentar apenas um *link* por *site*, dando ao usuário a opção de ver todos os demais *links* da resposta associados àquele *site*.

MOTORES DE BUSCA OU DIRETÓRIOS?

Conforme visto, existem diferenças essenciais entre motores de busca e diretórios, o que faz com que existam vantagens e desvantagens associadas ao uso de cada um dos tipos de ferramentas. Os diretórios têm bases de dados menores, mas que contêm informações mais relevantes. Por exemplo, ao se procurar, utilizando-se a árvore hierárquica de assuntos, o tópico "motores de busca" (*search engines*) no diretório Yahoo!, só se encontrarão itens relevantes. O mesmo não acontecerá, caso efetuemos uma pesquisa com a palavra-chave *search engines* em um motor de busca como o Altavista. Neste caso, obter-se-ia mais de um milhão e meio de resultados, e não há garantia de que os itens recuperados sejam relevantes. Diretórios são também mais apropriados para buscas por tópicos que sejam de interesse para um grande número de pessoas, pois é alta a probabilidade que sejam parte da árvore hierárquica; ou tópicos muito amplos os quais retornariam um número muito elevado de respostas em um motor de busca. Já os motores de busca permitem a localização de qualquer tipo de informação, por mais obscura ou específica, desde que exista na Internet e esteja indexada. Mas como a sua base de dados é muito grande, constituída de milhões de páginas, a chance de se recuperar um grande número de resultados não relacionados com os tópicos pesquisados é também maior. Ou seja, obtém-se menor precisão nos resultados da busca. Por outro lado, paradoxalmente, apesar de terem maiores bases de dados, as aranhas dos motores de busca podem não indexar alguns tipos páginas que poderiam ser incluídas nos diretórios (como, por exemplo, as informações que fazem parte da *Web* invisível).

Os motores de busca procuram compensar o excesso de itens recuperados com seus mecanismos internos de ordenação por relevância, mostrando em primeiro lugar os que, de acordo com seus critérios, deveriam ser os mais importantes. Uma vez obtida a lista dos resultados, o

usuário pode ler as descrições para decidir quais os *sites* serão de maior interesse. No caso dos diretórios, especialmente dos diretórios avaliados, esta descrição pode ser de melhor qualidade. As descrições dos motores de busca, por serem elaboradas automaticamente, podem não conter informações adequadas para facilitar a decisão do usuário. Os robôs não podem, por exemplo, identificar o tema central ou gênero literário de um documento e podem não detectar elementos importantes das páginas como gráficos ou imagens, assim como não podem extrair de um documento dados como o seu autor e sua afiliação institucional ou mesmo a data de publicação. Acessar o *site* pode ser a única maneira de verificar se os recursos são relevantes ou não.

Deve-se ter em mente, também, que, ao se pesquisar em um diretório, a consulta é feita apenas no título, categoria e uma breve descrição dos documentos. Já os motores de busca, em sua maioria, proporcionam uma pesquisa no texto integral dos documentos. Ou seja, o termo de busca poderá ser encontrado onde quer que seja que ele apareça no documento.

Outra diferença importante entre os motores de busca e diretórios é a rapidez com que a informação é incluída. Como nos diretórios, a inclusão de uma informação exige o trabalho humano de avaliação e seleção de recursos, uma página submetida a eles pode demorar pelo menos um mês para ser incluída. No caso dos motores de busca, que usam indexação automática, este tempo costuma ser mais rápido, e suas bases de dados contêm informações mais recentes.

Deve-se observar que, hoje em dia, a distinção entre motores de busca e diretórios já não é tão nítida e que a maioria deles pode ser considerada **ferramenta híbrida**. Os diretórios permitem buscas por palavras-chave em suas categorias, e os motores de busca, por sua vez, têm incluído diretórios em suas páginas principais. No caso dos diretórios, isso acontece, porque, mesmo sendo seletivos, o número de *sites* incluídos já é muito grande, dificultando aos usuários encontrar os itens procurados apenas através da navegação entre as categorias. O LookSmart, por exemplo, possui cerca de 60 mil subcategorias e indexa mais de um milhão de URLs. Além disso, os diretórios têm feito parcerias com motores de busca, para que, na eventualidade de um usuário não encontrar o que deseja, eles não recebam uma resposta negativa: nada encontrado. Nestes casos, automaticamente e de maneira transparente, o diretório aciona o motor de busca e traz da Internet *sites* que contenham as palavras-chave. Yahoo!, por exemplo, submete as palavras-chave dos usuários ao Inktomi. Caso não haja documentos em sua própria base de dados, a ferramenta retorna *sites* não incluídos pelo diretório.

Por outro lado, para proporcionar aos usuários uma opção de maior seletividade de recursos, os motores de busca têm feito parcerias com diretórios e incluído *links* selecionados em sua página principal. É possível encontrar um destes diretórios na página principal de quase todos os grandes motores estrangeiros. O Altavista e Excite, por exemplo, têm parceria com o diretório LookSmart, e o HotBot e Lycos, com o Open Directory.

METAMOTORES

Para se obterem resultados melhores em uma pesquisa de informação na *Web*, é recomendável que se utilizem várias ferramentas, já que, segundo alguns estudos, há pouca superposição na informação recuperada por motores diferentes. Para facilitar este processo, foram criados os metamotores (também chamados de multibuscadores), que permitem a execução de uma mesma busca em mais de uma ferramenta (motores ou diretórios), ao mesmo tempo exibindo todos os resultados encontrados em uma só lista. Estas ferramentas não possuem nenhuma base de dados, utilizando exclusivamente dados de outras ferramentas de busca. Exemplos de metamotores são Inference Find (http://www.infind.com), SavvySearch (http://www.savvysearch.com), Mamma (http://www.mamma.com), MetaMiner (http://miner.bol.com.br) e MetaBusca ZAZ (http://metabusca.zaz.com.br/busca/metabusca/home.htm), sendo estes dois últimos brasileiros.

Os metamotores fornecem uma interface que permite ao usuário formular a busca e clicar em um botão para receber os resultados da pesquisa. Geralmente, fazem um pré-processamento da consulta do usuário para prepará-la para submissão a cada ferramenta, e a maioria oferece processamento pós-busca para compilar os resultados. Algumas ferramentas que se intitulam metamotores são, na realidade, pseudometamotores, pois que apenas fornecem uma interface onde vários motores são listados sem que haja um mecanismo de busca integrada. Nestes casos, há uma caixa de pesquisa para cada motor, e as consultas são entradas e submetidas separadamente para cada ferramenta. Beaucoup Search Engines (http://www.beaucoup.com/engines.html) é um exemplo de uma ferramenta que funciona nestes moldes.

Existem alguns metamotores que funcionam através de um *software* instalado diretamente no microcomputador e que podem facilitar a construção local de estratégias de busca e conter muitas outras ferramentas de apoio que podem auxiliar, por exemplo, na eliminação de *links* duplicados ou mortos, armazenagem de buscas e ordenação dos resultados. Alguns exemplos destas são o *freeware* WebFerret (http://www.ferretsoft.com/netferret/), Mata Hari (http://www.thewebtools.com/), Copernic (http://www.copernic.com/) e BullsEye (http://www.intelliseek.com)

DIFERENÇAS ENTRE OS METAMOTORES

Como nos casos das ferramentas apresentadas anteriormente, existem variações entre os metamotores. Eles apresentam diferenças em relação à interface de busca, motores utilizados na pesquisa, modo de processamento das consultas, bem como forma de compilação e apresentação dos resultados.

Quanto à interface de pesquisa e aos motores utilizados, muitos fazem a busca em 6 a 10 motores, geralmente selecionados entre os maiores, como Altavista e HotBot. Outros oferecem mais opções: o SavvySearch, por exemplo, lista mais de uma centena de motores à escolha do usuário. Nestes casos, os metamotores podem funcionar como os diretórios de ferramentas temáticas, descritos anteriormente, permitindo que os usuários selecionem ferramentas especializadas em algum idioma, ou assunto. Algumas interfaces mostram as ferramentas utilizadas em listas facilmente visualizáveis e permitem que o usuário selecione em quais das ferramentas oferecidas quer pesquisar; outras podem não permitir esta personalização ou mesmo não indicar, nem mesmo em suas páginas de ajuda, quais motores são pesquisados. Os metamotores também podem efetuar buscas em outras partes da Internet como os arquivos de grupos de discussão da *Usenet* ou em *newswires*.

Quanto ao processamento da consulta, a maioria dos metamotores permite a formulação de uma expressão de busca em uma sintaxe semelhante à usada pela maioria dos motores, podendo permitir também o uso de lógica booleana e mesmo de linguagem natural. Alguns traduzem as consultas para a linguagem utilizada pelos motores individuais. Outros não o fazem, enviando a consulta como entrada pelo usuário, o que pode prejudicar a eficiência da busca, pois cada motor de busca usa uma sintaxe específica. Por exemplo, alguns motores de busca aceitam os conectivos booleanos (AND, OR, NOT), e outros aceitam apenas sinais de inclusão e exclusão (+ , -). Portanto, dependendo de como a consulta for repassada ao motor, ela pode não ser corretamente interpretada por este.

O tempo de resposta à consulta e o modo como os resultados são retornados são consideravelmente afetados pela forma em que as ferramentas são pesquisadas: seqüencial ou simultaneamente. É comum a interface permita que o usuário especifique um tempo limite de espera pelos resultados (por exemplo 10, 15 ou 30 segundos), acima do qual a busca seria cancelada para os motores que não apresentaram resultados. Alguns permitem também estabelecer o número de resultados a serem apresentados para cada motor pesquisado.

A forma mais recomendada de apresentação de resultados é aquela em que as respostas de cada ferramenta pesquisada são integradas, ordenadas por relevância (mostrando também quais motores retornaram resultados), e com resultados duplicados (ou seja, trazidos por mais de uma ferramenta) eliminados. Entretanto, algumas vezes os resultados produzidos por ferramenta pesquisada são agrupados e trazidos seqüencialmente. Alguns poucos, como o Inference Find, agregam os resultados por categorias. As listas de resultados podem conter apenas títulos ou mostrar títulos e curtas descrições, além do URL. Alguns podem mostrar a ordem de relevância que cada resultado obteve no motor que o recuperou (por exemplo, número 5 no Altavista, ou número 10, no HotBot).

Metamotores são indicados nos casos em que não se encontram muitos resultados quando se pesquisa um só motor. Podem também ser utilizados para verificar quais motores individuais trazem as melhores respostas e fornecer uma visão geral do que cada ferramenta contém sobre um tópico com fins de seleção de um motor específico para uma busca expandida.

É importante notar que existem desvantagens com relação ao uso dos metamotores. A maior limitação é que os recursos de busca específicos de cada motor, que são os mecanismos para maior refinamento das pesquisas, tornam-se inacessíveis na interface do metamotor. Devido ao grande volume de informações na Internet, nos resultados obtidos, normalmente ocorre um nivelamento por baixo, ou seja, obtém-se maior quantidade de informações sem um correspondente aumento de qualidade. Por causa desta limitação, os metamotores são mais indicados para buscas onde se utilizam termos únicos ou outras buscas simples, que não requeiram maior sofisticação. Em alguns metamotores, apenas um subconjunto dos resultados de cada ferramenta

(geralmente os primeiros e, supostamente, mais relevantes) sejam recuperados. Buscas no metamotores tomam mais tempo porque processamento adicional é necessário para compilar os resultados e porque o tempo de resposta final será aquele da ferramenta mais lenta.

Como se manter atualizado sobre motores de busca

Como visto, as ferramentas de busca na Internet constituem um universo complexo, não só pelas diferentes características que apresentam individualmente, mas também pela variedade de tipos e subtipos e por estarem em constante evolução. Além disso, a dificuldade de se encontrarem informações relevantes através delas é mascarada por suas interfaces aparentemente amigáveis. Assim, apesar da grande quantidade de informações na *Web* e das ferramentas disponíveis para pesquisá-las, o usuário fica freqüentemente frustrado com os insatisfatórios resultados encontrados.

O profissional da informação deveria, minimamente, consultar a documentação, ainda que esta seja mais limitada que o desejável, de cada ferramenta, para melhor utilizá-la. Idealmente deveria se informar mais profundamente e se manter atualizado sobre elas. Existem *sites* na Internet que regularmente publicam artigos sobre as ferramentas de busca na Internet e tabelas comparativas de características dos motores. A seguir, listam-se alguns exemplos destes:

Search Engine Watch
(http://www.searchenginewatch.com)

SearchIQ (http:// www.searchiq.com/)

Search Engine Showdown
(http://www.searchengineshowdown.com/)

About.com Web Search Guide
(http://Websearch.about.com/)

Recomenda-se também a revista *Online*, que, além de sua versão impressa, disponibiliza alguns dos artigos publicados no URL http://www.onlineinc.com. Além do já citado *site* da biblioteca da Universidade de Berkeley, merece destaque o *site* mantido por Laura Cohen na University at Albany Libraries (http://www.albany.edu/library/internet/searchnet.html). O *site* Ferramentas de Busca na Internet (http://www.eb.ufmg.br/cendon/links/motores.htm) traz uma lista categorizada de ferramentas de busca.

REFERÊNCIAS BIBLIOGRÁFICAS

BLATTMANN, Ursula, FACHIN, Gleisy R. B, RADOS, Gregório J. Varvakis. *Recuperar a informação eletrônica pela Internet*. [*online*]. Disponível na Internet via WWW. URL: www.ced.ufsc.br/~ursula/papers/buscanet.html. Arquivo capturado em 08/06/2000.

GARMAN, Nancy, Meta search engines. *Online*, v. 23, n.3, p. 75-78, May/June 1999.

HAHN, Trudi Bellardo. Text retrieval online: historical perspective on Web Search Engines. *Bulletin of the American Society for Information Science*, v. 24,n. 4 ,7-10, April/May 1998.

HOCK, Randolph. Web search engines: features and commands. *Online*, v.23, n.3, p. 24-28, May/June 1999.

KIMMEL, Stacey. WWW search tools in reference services. *The reference librarian*, v.57, p.5-20, 1997.

LIMA, Cynthia Moreira. *O que é a Internet e como utilizá-la para pesquisa?* [*online*]. Disponível na Internet via WWW. URL: http://www.elo.com.br/~cynthia/interpesq.htm. Arquivo capturado em 09/06/2000.

POULTER, Alan. The design of World Wide Web search engines: a critical review. Program, v. 31, n. 2, p. 131-145, Apr 1997.

SCHWARTZ. Candy; Web search engines. *Journal of the American Society for Information Science*, v. 49, n.11, p.973-982, 1998.

SULLIVAN, Danny (Ed.). Search engine watch: tips about Internet search engine. [*online*]. Disponível na Internet via WWW. URL: http://www.searchenginewatch.com. Arquivo capturado em 09/06/2000.

SHERMAN, Chris. The future of Web search. *Online*, v. 23, n.3, p. 54-61, May/June 1999.

SULLIVAN, Danny. Crawling under the hood: an update on search engine technology, *Online*, v. 23, n.3, p. 30-38, May/June 1999.