UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE BIBLIOTECONOMIA E COMUNICAÇÃO
Departamento de Ciências da Informação
Curso de Biblioteconomia

Juliana Hugo

SISTEMAS DE BUSCA NA WORLD WIDE WEB

Porto Alegre

2005

Juliana Hugo

SISTEMAS DE BUSCA NA WORLD WIDE WEB

Monografia apresentada como requisito parcial à obtenção do grau de Bacharel em Biblioteconomia, sob a orientação do Prof. Dr. Rafael Port da Rocha, do Curso de Biblioteconomia, Departamento de Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul.

Porto Alegre

2005

H892s Hugo, Juliana

Sistemas de busca na world wide web [manuscrito] / Juliana Hugo; Orientação [por] Rafael Port da Rocha. – Porto Alegre, 2005.
75 fls.

1. Internet 2. World Wide Web 3. Recuperação da Informação 4. Sistemas de busca 5. Mecanismos de busca 6. Diretórios 7. Metabuscadores I. Rocha, Rafael Port da II. Título.

CDD 025.4

Departamento de Ciências da Informação
Rua Ramiro Barcellos, 2705.
CEP: 90035-007
Tel: (51) 33165146
Fax: (51) 33165435

A minha mãe, bibliotecária, que me deu o exemplo de uma profissão e ao meu pai, que me fez ver muito além do visível e que me ajudou na decisão de escolher essa profissão. A toda minha família que me deu o suporte e incentivo que necessitei, principalmente a minha filha Camila pela paciência e pela compreensão da nossa "distância" durante a realização do curso.

Ao meu marido Marcos, pela força e coragem que possui. Pelo amor e respeito que me tem dedicado todos esses anos, pelas palavras de incentivo e carinho que me deu durante os tempos de faculdade e que ainda me dá.

Aos amigos feitos ao longo do curso pelos bons momentos que passamos. Em especial a querida amiga e Bibliotecária Rosângela Martins de Arruda que sempre esteve ao meu lado, nos bons e maus momentos, demonstrando sua amizade e sinceridade.

Ao Prof. Doutor Rafael Port da Rocha pela dedicação e sabedoria com que auxiliou a elaboração deste trabalho, dando dicas preciosas nos momentos oportunos.

A todos os que fizeram parte da minha formação profissional. Em especial as bibliotecárias, Rosária Geremia, e Irma Macolmes que se tornaram além de colegas, grandes amigas.

A cada dia, nos tornamos mais e mais dependentes das facilidades e serviços providos pela Internet. A existência de ferramentas de busca como Google, por exemplo, nos permite realizar pesquisas e encontrar respostas para nossas dúvidas sobre inúmeros assuntos no conforto de nossos lares. Se podemos ter respostas para nossas perguntas na ponta dos dedos, porque ir a biblioteca!!??? Muito se ouve falar sobre os dias contados desses "repositórios de livros" em função da disseminação do uso da Internet. Mas será que isso vai mesmo acontecer ? Usando uma forma um tanto irônica, o autor diz que isso não ocorrerá, e tenho que concordar. Devemos usufruir ao máximo a facilidade proporcionada pela Web, mas lembrar que as bibliotecas mudaram muito e que ainda são indispensáveis para nossa educação pessoal e profissional.

Mark Herring

RESUMO

O grande crescimento na quantidade de informações disponíveis na Internet originou a necessidade de criação e de uso de sistemas de recuperação de informação na World Wide Web. Atualmente encontramos na web uma enorme quantidade de sistemas de busca, cada um com suas peculiaridades e características. Este trabalho reflete sobre algumas diferenças existentes entre a Internet e a World Wide Web. Oferece uma visão geral dos sistemas de busca, expressão que abrange diretórios, mecanismos de busca e metabuscadores. Descreve suas características, diferenças e semelhanças, analisando e comparando através de experimentos as vantagens e desvantagens do uso dos diferentes sistemas e suas ferramentas com o objetivo de proporcionar aos profissionais da informação subsídios para um aumento na qualidade da recuperação de informações relevantes através da World Wide Web.

Palavras-chave: Internet. World Wide Web. Recuperação da Informação. Sistemas de busca. Mecanismos de busca. Diretórios. Metabuscadores.

ABSTRACT

The growth of information amount available in Internet originated the necessity to create and use information retrieval systems. Today, we have a large amount of search engines in the internet, each one with its own peculiarities and characteristics. This work shows some differences between Internet and web and offers an overview about the search engines, a term that involves the directories, search mechanisms and metasearchers. This work describes characteristics, differences and similarities among search engines tools, analyzing and comparing the advantages and disadvantages of the use of these different systems. It also suggests search tips to be used in each different system, with the goal to provide to professional of the information a basis to increase the quality of relevant information retrieval through the World Wide Web.

Keywords: Internet; World Wide Web; Retrieval information; Search Tools; Motors of Search; Directories; Metasearchers.

LISTA DE FIGURAS, TABELAS E GRÁFICOS

p.

SUMÁRIO

1 INTRODUÇÃO

A quantidade de informações disponíveis na internet cresce diariamente. Para a recuperação dessas informações, que muitas vezes, não estão disponíveis em bases de dados especializadas, são utilizados sistemas de busca na internet. A utilização da internet como meio de recuperar informações é importante, pois possui algumas vantagens. A informação é extremamente atualizada, o acesso é fácil, ela funciona 24 horas ao dia e acessa uma grande variedade de recursos. As desvantagens estão ligadas muitas vezes, a falta de critérios para avaliar a qualidade dos resultados; a falta de organização da web como um todo e também a mecanismos de busca utilizados de forma errada ou sem uma utilização adequada de todas as suas ferramentas. A dificuldade dos profissionais da informação está, muitas vezes, justamente em não conhecer os recursos disponíveis pelos sistemas de busca da web, o que faz com que os resultados das buscas realizadas através desses sistemas sejam irrelevantes.

O tema do presente trabalho é um estudo sobre sistemas de busca na web, envolvendo diretórios, mecanismos de busca e metabuscadores. Desta forma, tem como objetivo principal, identificar as características, diferenças e semelhanças entre esses sistemas, com o intuito de oferecer subsídios aos profissionais da informação para que os mesmos obtenham resultados mais eficientes nas suas pesquisas bibliográficas. Para alcançar o objetivo acima proposto, foi realizada uma revisão bibliográfica na literatura científica envolvendo as áreas da Biblioteconomia e Ciência da Computação, nas quais foram identificados os conceitos que envolvem os sistemas de busca, seu funcionamento e sua organização.

Este trabalho está estruturado da seguinte forma:

√ Revisão da Literatura. Traz conceitos de vários autores sobre o tema para adquirir

subsídios para o estudo. Expões alguns conceitos, características e diferenças entre a Internet e a World Wide Web. Trata sobre a recuperação da informação na web. Conceitua os sistemas de busca, considerando que a expressão abrange três categorias: os mecanismos de busca, os diretórios e os metabuscadores. Os mecanismos de busca são os sistemas baseados no uso exclusivo de programas de computador para a indexação das páginas da web, os diretórios são os sistemas de busca nos quais a indexação das páginas da web é realizada por pessoas e os metabuscadores são programas de computador que localizam informações em diversos sistemas de busca simultaneamente.

√ O capítulo 3, apresenta a metodologia utilizada.

√ O capítulo 4, Apresentação e Análise dos Dados, contém todas as informações colhidas na investigação bibliográfica. Estabelece comparações entre os diferentes sistemas de busca, e relata os resultados obtidos na experimentação feita através de buscas entre os diferentes sistemas.

√ O capítulo 5, Conclusão, traz a análise final de todos os dados e impressões pessoais sobre o trabalho.

1.1 Justificativa

A busca faz parte da rotina do profissional da informação. Mais do que tudo, a recuperação da informação é um dos objetivos principais nesta área. E o que fazer quando se esgotam os recursos disponíveis e a informação não é encontrada?

Mesmo os artigos, periódicos e livros on-line precisam de um tempo para publicação na web. Mas além desses recursos, a web dispõe de sites pessoais e institucionais onde seus autores disponibilizam seus trabalhos em questão de minutos. É verdade que nem tudo que é publicado na web é confiável, mas a recíproca também é

verdadeira. É por isso que o conhecimento acerca dos sistemas de busca é fundamental na rotina do profissional da informação.

Cresce a cada dia a necessidade de o profissional da informação conhecer as mais variadas fontes bibliográficas, isso porque a informação também faz parte da globalização dos povos. A cada dia, e com mais freqüência, localizamos materiais de uma área específica do conhecimento em bases de dados especializadas de outras áreas. É a interdisciplinaridade que tanto faz parte do nosso contexto informacional.

Para que o profissional da informação esteja preparado para responder qualquer tipo de pergunta vinda do seu usuário ele precisa, antes de tudo, saber como recuperar a informação. Quando esta informação esta disponível na base de dados que o profissional costuma utilizar, o sucesso é garantido. No entanto, existe um grande número de informações à margem das bases de dados especializadas. Para acessá-las, o uso de sistemas de busca na web é indispensável. Neste contexto questiona-se se o profissional da informação possui o conhecimento necessário/adequado/suficiente acerca da utilização dos sistemas de busca existentes, assim como de suas ferramentas ou recursos.

Um estudo apresentado por Ivonen (1995 apud DETERS; ADAIME, 2003, p. 1), demonstra que aqueles que conhecem o funcionamento interno de um mecanismo de busca e possuem experiências com a linguagem de consulta tem mais probabilidade de encontrar a informação desejada.

Por isso, este trabalho visa, primeiramente, obter maiores conhecimentos sobre o tema tratado, por tratar-se de um assunto relativamente novo e com grandes interesses comerciais. Muitas vezes, o sigilo por traz destes interesses comerciais impedem um conhecimento aprofundado sobre muitos mecanismos e, devido a esses mesmos fatores, existe pouca literatura publicada, principalmente no idioma

português. Este trabalho visa também que, através deste conhecimento, a busca seja aprimorada para uma maior eficiência na pesquisa bibliográfica realizada pelos profissionais da informação.

1.2 Limitação do Estudo

Este estudo limita-se a descrever e explorar os sistemas de busca de propósito geral. Os sistemas de busca temáticos ou especializados em qualquer área do conhecimento ou pertencentes a qualquer órgão ou instituição não pertencem ao universo de estudo desta monografia.

Com relação a Web, as buscas questionadas neste trabalho, feitas através de sistemas de busca, são pertencentes a web “aberta”. A web invisível ou profunda e/ou qualquer tipo de sistema de busca ou site pago ou que possua algum tipo de senha de acesso não pertence ao universo deste estudo. Mesmo porque os sistemas de busca não são capazes de indexar este tipo de documento. Isso ocorre porque os sites da web invisível não apresentam a informação no texto corrido de uma página HTML. O seu conteúdo encontra-se oculto dentro de bases de dados. Só a título de curiosidade, a Web Invisível pode ter 500 vezes a dimensão da Web Visível. (ROSAS; HOURMANT, 2001).

2 REVISÃO DA LITERATURA

2.1 Internet x World Wide Web

Existe uma certa confusão, por parte dos usuários da internet, em relação à Internet e a *World Wide Web*. Muitos utilizam os dois conceitos como se fossem sinônimos. Isto não é verdade, a *World Wide Web* é apenas uma pequena parte integrante da Internet. Para compreender melhor estes dois conceitos vejamos algumas definições e um pouco de história.

A internet é "o conjunto de diversas redes de computadores que se comunicam através dos protocolos TCP/IP" (CASTRO, 2003, p. 1), ou, segundo a resolução aprovada pela *Federal Networking Council norte-americano* em 1995 que definiu o termo internet em consulta a membros da Internet e comunidades de direitos da propriedade intelectual que diz o seguinte:

> Internet se refere ao sistema de informação global que -- (i) é logicamente ligado por um endereço único global baseado no Internet Protocol (IP) ou suas subseqüentes extensões; (ii) é capaz de suportar comunicações usando o Transmission Control Protocol/Internet Protocol (TCP/IP) ou suas subseqüentes extensões e/ou outros protocolos compatíveis ao IP; e (iii) provê, usa ou torna acessível, tanto publicamente como privadamente, serviços de mais alto nível produzidos na infra-estrutura descrita.

Em 1982 pela primeira vez, foi registrado o uso da palavra Internet. Esse conceito tinha como fundamento não mais uma rede de computadores interligados, mas várias redes se intercomunicando. Foi criado o protocolo TCP/IP, a linguagem comum de todos os computadores da Internet. Em 1984 o número de servidores na Internet era superior a mil. Foi o ano também, do primeiro sistema operacional usado no círculo restrito de hackers, o DOS 3.0. Em 1987 o número de servidores

da Internet superava os 10.000. Em 1990 os servidores ultrapassavam os 300.000. “The World Comes Online” (www.std.com) é o primeiro provedor privado de acesso discado. Foi construída a HTML, a linguagem em que se escrevem páginas para serem vistas na rede. Em 1991 além de serem lançados o Windows 3.1, o Quicktime, os formatos JPEG/MPEG, abriu-se a primeira conexão do Brasil com a Internet por meio da Fapesp e foi criada a World Wide Web.

Em 1994 a quantidade de informações na internet já era imensa. Essa quantidade fez com que se criasse um índice geral. Foi criado então o primeiro diretório, o Yahoo!, que permitia buscas de páginas apenas por categoria (hoje a busca pode ser feita também através de palavras-chaves). Em 1995, Compuserve, America On-Line (AOL) e Prodigy começaram a oferecer acesso discado comercial. Em 2000 a Internet em banda larga começou a operar comercialmente no Brasil usando linhas telefônicas e cabeamento para TV.

A World Wide Web, chamada popularmente de Web, nasceu em 1991, no laboratório CERN, na suíça. Seu criador foi Tim Berners-Lee. Ele a desenvolveu como uma linguagem que serviria para interligar computadores do laboratório e outras instituições de pesquisa e exibir documentos científicos de forma simples e de fácil acesso. Para entender o que é World Wide Web (também chamada Web ou WWW) será utilizada a definição de CASTRO (2003, p. 1):

> [...] é, em termos gerais, a interface gráfica da Internet. Ela é um sistema de informações organizado de maneira a englobar todos os outros sistemas de informação disponíveis na Internet. Sua idéia básica é criar um mundo de informações sem fronteiras, prevendo as seguintes características... interface consistente; incorporação de um vasto conjunto de tecnologias e tipos de documentos; "leitura universal". Para isso, implementa três ferramentas importantes... um protocolo de transmissão de dados - HTTP; um sistema de endereçamento próprio - URL; uma linguagem de marcação, para transmitir documentos formatados através da rede - HTML.

Hoje a World Wide Web é o segmento da Internet que mais cresce. Só para se ter uma idéia do tamanho da Web, o Google, um dos maiores sistemas de busca da Web, cabe resaltar que ele não indexa todo o conteúdo da web, indexa aproximadamente 560 milhões de documentos (veja Tabela 1). Na Web, que surgiu como uma ferramenta de uso acadêmico, podemos encontrar informação sobre praticamente qualquer assunto, como, instituições públicas, órgãos de governo, organizações comerciais, notícias, educação (inclusive cursos de ensino à distância), jogos online, bibliotecas, universidades, órgãos públicos, empresas, pessoas, clipes de vídeo, arquivos de som, animações interativas, arquivos para download, programas, Design, Hipertexto, Multimídia e Textos.

Como pode-se concluir, assim como as informações disponíveis, a internet e a *World Wide Web* vem crescendo cada vez mais e mais rápido. Hoje existem até salas de bate-papo onde autores discutem e opinam sobre o futuro dos avanços na internet e suas perspectivas. Para ilustrar a história e a evolução da internet, vejamos a figura 1.

FIGURA 1 - Evolução da Internet e Sites de Busca

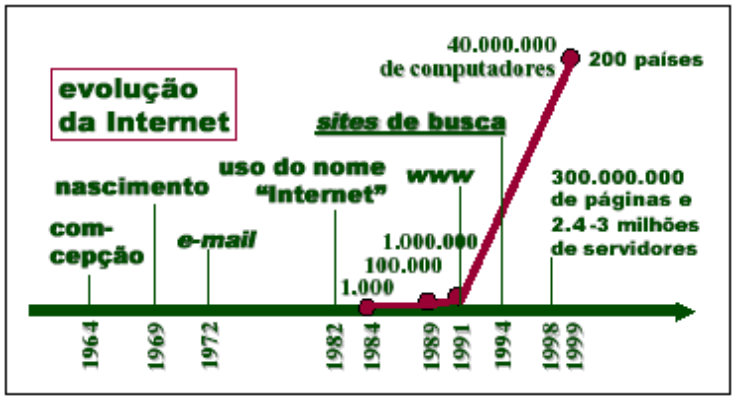


Figura retirada do trabalho de GONZALEZ (2002, p. 2).
Disponível em: < http://www.inf.pucrs.br/~gonzalez/docs/qualificacao.pdf.>
Acesso em: 03 de ago. 2005.

2.2 Recuperação da Informação (RI)

Mesmo antes da utilização da internet e de computadores em unidades de informação, a recuperação da informação já era objeto de estudo dentro das ciências da informação. Com a inserção da internet e, conseqüentemente, com o aumento da quantidade das informações produzidas e a facilidade e rapidez com que eram publicadas, surgiu também um aumento na dificuldade de recuperar informações em sistemas manuais. Isso porque não era possível promover uma organização tão rápida ao acervo, quanto a atualização que recebia a internet. Para a maioria dos profissionais da informação, como relata MARCONDES, (s. d., p. 1), “ [...] a perspectiva de um uso intenso das tecnologias da informação é associada a uma valorização de suas atividades técnicas”. Embora muitos profissionais não concordem com este ponto de vista, a internet proporciona uma maior visibilidade para o profissional da informação que pode fazer com que sua instituição e os serviços oferecidos pela mesma, sejam vistos em várias partes do mundo.

Segundo DETERS e ADAIME (2003, p. 1), a internet é uma das principais fontes de informação para muitos usuários na atualidade. Isso, justamente pela rapidez com que é atualizada e pela facilidade de acesso que proporciona.

> As ferramentas de busca são os meios mais usados [...] para encontrar novos Web sites on-line, usadas por 73,4% daqueles entrevistados [...]. Do total, 84,8% das pessoas usam ferramentas de busca para encontrar novos web sites [...]. (SILVEIRA, 2002, p. 22)

O processo de recuperação da informação compreende **indexar** (representação**), armazenar** e **recuperar** (acesso) os itens de informação. Seu principal objetivo é a seleção dos documentos relevantes a uma necessidade de informação, em meio a uma quantidade maior de documentos tanto relevantes,

quanto irrelevantes. Segundo TEIXEIRA e SCHIEL (1997, p. 1) esses três processos (indexar, armazenar e recuperar) tornaram-se mais fáceis com a inserção da informática nas unidades de informação. O computador foi, inicialmente, utilizado pelos profissionais da informação para armazenar informações bibliográficas e gerar índices impressos. Atualmente os sistemas informatizados oferecem uma infinidade de recursos de busca, como o uso dos operadores booleanos, por exemplo. As unidades de informação, pelo menos grande parte delas, visando a recuperação da informação, já possuem suas bases de dados que podem ser armazenadas em meios magnéticos ou ópticos e acessados, local ou remotamente.

Com a inclusão da informática nos processos de recuperação da informação, o acesso à informação se tornou mais democrático. Porém, a falta de padronização na linguagem utilizada trouxe dificuldades para os usuários destes sistemas. Muitas vezes, a necessidade de informações do usuário não pode ser descrita diretamente. Cabe ao profissional da informação transformar essa necessidade em consulta válida, ou seja, que essa consulta possa ser processada por um sistema de busca, por meio de palavras-chaves.

Segundo Wives (2002 apud DETERS; ADAIME, 2003) o usuário precisa traduzir a sua necessidade de informação, utilizada em linguagem formal específica de um sistema de RI, o que, segundo ele, representa uma das maiores dificuldades para o usuário. Macedo (2001 apud DETERS; ADAIME, 2003) conclui que esta dificuldade do usuário na formulação da necessidade de informação acontece porque o usuário possui uma “necessidade visceral”, ou seja, ele está consciente de sua necessidade, no entanto, não consegue uma definição, sequer em linguagem natural dessa necessidade. O que torna mais difícil a tarefa de traduzir essa linguagem natural em uma linguagem capaz de ser compreendida pelo sistema automático de RI. Wives (2002 apud DETERS; ADAIME, 2003) também chama

atenção para este problema e coloca que um documento pode ser relevante à consulta do usuário mas pode não ser relevante para o usuário que por sua vez pode ter formulado incorretamente a sua necessidade de informação.

> [...] no âmbito da recuperação da informação, a estratégia de busca pode ser definida como uma técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados." (LOPES, 2002, p. 61).

2.3 Sistemas de Busca na World Wide Web

A Internet permite que pessoas de diferentes culturas, conhecimentos e interesses possam compartilhar informações, tornando disponíveis suas informações publicamente e, ao mesmo tempo, procurar por conhecimentos e experiências que outros desenvolveram. As informações disponíveis na Internet podem estar disponíveis em forma de textos, arquivos formatados, imagens, sons, vídeos, etc.

Os sistemas de busca surgem como uma tentativa de recuperar a informação contida na internet, isto é, como uma tentativa de colocar ordem em um grande caos. Seu objetivo é encontrar informação do interesse do usuário na World Wide Web. Segundo VIDOTTI (2004, p. 3), o aumento na quantidade de informações disponíveis na internet deve-se, em grande parte, pelo surgimento de ferramentas que permitem a construção, de forma rápida e fácil, de páginas e sites. A mesma pessoa que escreve um artigo pode criar um site para publicar seu artigo. O que faz com que não esqueçamos do fator relativo a confiabilidade dos documentos encontrados na rede e que olhemos sempre com olhos críticos as informações e o contexto obtido em um resultado de busca. Devemos lembrar também que os sistemas de busca estão, hoje em dia, faturando milhões de dólares com anúncios e

propaganda, o que pode influenciar nos resultados apresentados.

> Procurar uma informação na internet assemelha-se, em muitos aspectos, a procurar uma informação em uma biblioteca do tipo que admite o acesso do público às estantes onde estão guardados os livros. No caso deste tipo de biblioteca, você pode optar por três atitudes. A primeira é recorrer ao funcionário encarregado de atender o público, e pedir que o ajude a localizar tal obra, ou tal assunto. Qualquer que seja o modo como a biblioteca esteja organizada, é necessário que você saiba o que quer, e que possa comunicar ao funcionário, com clareza, o que deseja encontrar. Este é o único caso em que a internet não segue a biblioteca: não existem funcionários para facilitar a vida dos internautas. Mas você vai ver que os mecanismos de busca tentam suprir essa ausência. (MOTA, 1998, p. 16).

Estes sistemas de busca na Web podem ser definidos segundo suas categorias, são elas, mecanismos de busca, diretórios e metabuscadores. Esses três tipos de sistemas possuem diferenças e particularidades. Suas semelhanças estão justamente no objetivo em comum, recuperar a informação existente na *web* e torna-la acessível ao usuário. Embora existam representantes, tanto de mecanismos e diretórios como de metabuscadores, dirigidos a uma área específica do conhecimento, ou mais de uma, esses não irão fazer parte deste trabalho. Neste trabalho são abordados apenas os sistemas de busca de propósito geral.

2.4 Mecanismos de Busca

Conhecidos como mecanismos de busca, motores de busca ou ainda máquinas de busca. Este tipo de sistema não foi o primeiro a surgir, no entanto, podemos considerá-lo como o mais popular entre os sistemas, um grande exemplo dessa popularidade é o motor *Google* que a cada dia recebe um maior destaque.

> Um motor de busca é uma espécie de catálogo mágico. Mas, diferente dos livros de referência comuns, nos quais está acessível à informação que alguém organizou e registrou, o catálogo do motor

> de busca está em branco, como um livro vazio. Ao se realizar uma consulta, a lista de ocorrências de assunto é criada em poucos segundos por meio do trabalho de um conjunto de softwares de computador conhecidos como *spiders* (aranhas), que vasculham toda a web em busca das ocorrências de um determinado assunto em uma página. Ao encontrar uma página com muitos *links*, os *spiders* embrenham-se por eles, conseguindo, inclusive, vasculhar os diretórios internos – desde que eles sejam públicos, ou seja, tenham permissão de leitura para usuários – dos sites nos quais estão trabalhando. Motores de busca muito refinados são capazes de saber exatamente que atualizações houve em um site usando esse método de scanner. (SEGREDOS, 2004, p. 7)

Os mecanismos de busca, ou motores de busca, possuem três componentes principais, um programa de computador denominado **robô** (*robot, spider, crawler, wanderer, knowbot, worm* ou *web-bot)*. Esse robô visita os *sites* ou páginas armazenadas na web e , ao chegar em cada *site*, o programa robô cria uma cópia ou réplica do texto contido na página visitada e guarda essa cópia em sua base de dados.

O segundo componente é uma **base de dados** constituída das cópias efetuadas pelo robô. Essa base de dados, também denominada índice ou catálogo, fica armazenada no computador, também chamado servidor do mecanismo de busca.

O terceiro componente é o **programa de busca** propriamente dito. Esse programa de busca é acionado cada vez que alguém realiza uma pesquisa. Nesse instante, o programa percorre a base de dados do mecanismo em busca dos endereços - os URL - das páginas que contém as palavras, expressões ou frases informadas na consulta. Em seguida, os endereços encontrados são apresentados ao usuário, agrupados, em uma lista.

Esses três componentes estão associados às três funções básicas de um sistema de busca, a indexação ou "cópia" das páginas da web; o armazenamento das "cópias" efetuadas (no google a cópia é representada pela expressão “em

cache”) e a recuperação das páginas que preenchem os requisitos indicados pelo usuário por ocasião da consulta. Vale destacar que,

> [...] ao realizar uma pesquisa, quer seja através de um mecanismo de busca quer seja através de um diretório, você não está pesquisando diretamente a web. Você está pesquisando uma base de dados localizada num site da web. Nessa base de dados, encontra-se uma cópia, uma fotografia ou, fazendo uma analogia, às vezes uma simples foto 3 x 4 dos sites e páginas existentes na web. (MOURA, 2001, p. 6).

Para criar a base de dados de um mecanismo de busca, o programa robô sai visitando os *sites* da *web*. Ao passar pelas páginas de cada site, o robô anota os URL existentes nelas, para depois ir visitar cada um desses URL. Visitar as páginas, fazer as cópias e repetir a mesma operação, cópia e armazenamento, na base de dados, do que ele encontrar nesses *sites*. Essa é uma das formas de um mecanismo de busca encontrar os *sites* na *web, a* outra maneira é o criador ou responsável pelo *site* informar, ao mecanismo de busca, qual o endereço, o URL, do *site*. Todos os mecanismos de buscas têm, geralmente, um quadro reservado para o cadastramento, submissão ou inscrição de novas páginas. É um *hiperlink* que recebe diversas denominações conforme o sistema de busca.

2.4.1 Tamanho das Bases de Dados

Os mecanismos de busca, sempre quando objetos de estudo em relação a sua qualidade, tem seus tamanhos divulgados. Isso porque esta é uma das únicas maneiras concretas de avaliarmos, de certa forma, sua qualidade. Quanto maior a base de dados do mecanismo de busca, maior é também a chance dele recuperar a informação procurada. “[...] os motores maiores tendem a ser mais usados, atraindo maior número de anunciantes e podendo cobrar maiores taxas pelos anúncios”

(CÉNDON, 2001, p. 42). Através da tabela abaixo, podemos perceber que o Google esta no lugar do Altavista, que foi considerado por muitos anos como o maior mecanismo de busca do mundo. Devemos notar, entretanto, que mesmo ele, indexa apenas 56% da World Wide Web, ou seja, através dele podemos ter acesso a praticamente metade dos documentos existentes na rede.

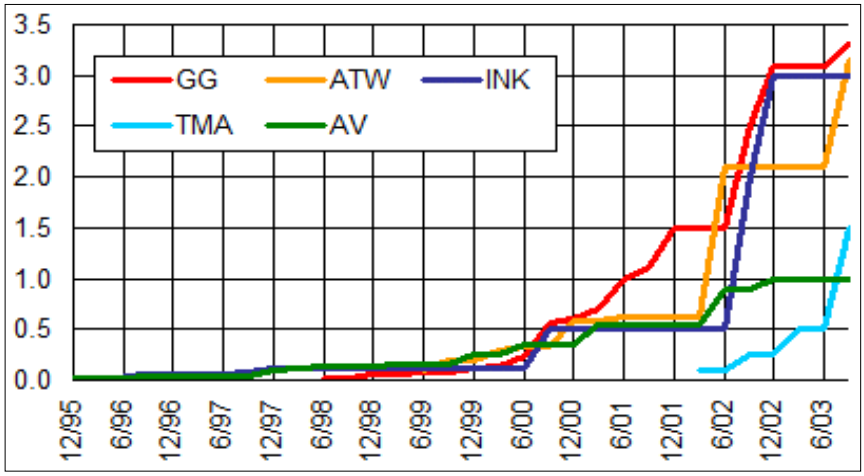
Tabela 1 - *Tamanho da base de dados dos motores de busca

Motor de busca	Nº de páginas (em milhões)	% da Web
Google	560	56,00%
WebTop.com	500	50,00%
Altavista	350	35,00%
Fast	340	34,00%
Northern Light	265	27,00%
Excite	250	25,00%
HotBot / Inktomi	110	11,00%
Go / Infoseek	50	5,00%
Lycos	50	5,00%

*Tabela criada pelo site *Search Engine Watc*h e retirada do texto de CENDÓN, 2001, p. 42. Disponível em: http://www.ibict.br/cienciadainformacao/viewissue.php?id=17. Acesso em 10 de maio de 2005.

Para se ter uma idéia do crescimento diário nas bases de dados dos mecanismos de busca, podemos analisar o gráfico 1, ele mostra a quantidade de documentos textuais indexados (em bilhões) para cada motor de busca no período de dezembro de 1995 até junho de 2003.

GRÁFICO 1 - Documentos Textuais Indexados



Legenda:

GG= Google
ATW= AllTheWeb
INK= Inktomi
TMA=Teoma
AV= AltaVista

Fonte: SEARCH ENGINE WATCH. *Search engine sizes*. Disponível em: <http://searchenginewatch.com/reports/article.php/2156481#key?>. Acesso em: 20 maio de 2005.

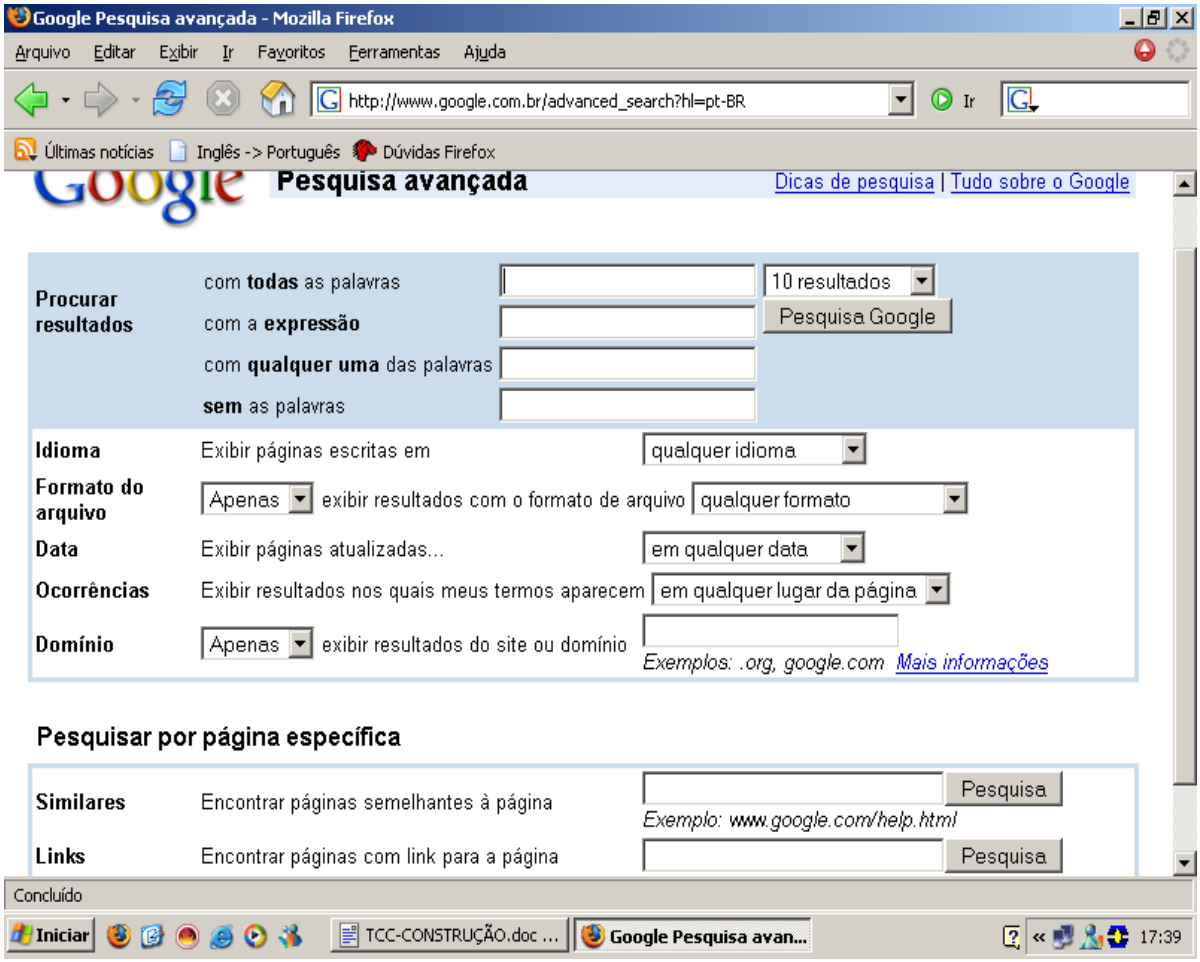
2.4.2 Operadores Booleanos

Hoje, para que se tenha sucesso na recuperação de informação através da Web é necessário, na maioria dos casos, que se faça uma “pesquisa avançada”. Ou seja, que se utilize ferramentas disponíveis no próprio sistema de busca (a maioria dos mecanismos de busca possui esta ferramenta) para delimitar a pesquisa. Essa delimitação pode ocorrer quanto ao idioma dos documentos, formato do arquivo, data, domínio, ocorrência (estabelece em que lugar da página o meu termo de consulta deverá aparecer), tamanho de imagens, etc. dos documentos recuperados.

> A forma mais usual de relacionamento entre termos é feita através do uso de operadores booleanos (e, ou, não). Esta é uma característica presente em quase todos os mecanismos de busca, geralmente sob o rótulo de “busca avançada”. Um problema comum é que às vezes o relacionamento é automático ou implícito, e nem sempre é fácil para o usuário identificar o operador booleano que é considerado quando digita apenas os termos, sem utilizar os conectores, ou seja, a operação default. (ALENCAR, 2001, p. 42)

Veja um exemplo de pesquisa avançada na figura abaixo, retirada do mecanismo de busca Google (www.google.com.br):

FIGURA 2 – Pesquisa Avançada



Disponível em: <Fonte: http://www.google.com.br/advanced_search?hl=pt-BR>
Acesso em: 12 maio 2005.

Além da utilização das ferramentas existentes nos próprios sistemas de busca e com vista a uma maior eficiência em nossas pesquisas, devemos utilizar também os operadores booleanos. Esse tipo de busca funciona com a inserção de símbolos que nos permitem fazer combinações, utilizando palavras, frases e expressões na pesquisa bibliográfica. Os operadores lógicos mais comuns, usados para relacionar termos ou palavras, na montagem de uma expressão de busca informatizada são, “AND”, “OR” e “NOT”, porém alguns sistemas de busca utilizam símbolos diferentes, outros trazem de maneira implícita os operadores booleanos. Veja na tabela abaixo quando e como eles devem ser utilizados.

TABELA 2 – Operadores Booleanos

Operador	Símbolo	Função	Exemplo
AND	&	Liga dois ou mais termos, limitando a busca. Somente as páginas que contêm todos os termos listados darão um bom resultado. (Restringe).	ciências & matemática mostrará páginas com ambos os termos: ciências e matemática.
OR	|	Liga dois termos e reúne todos os documentos que incluam pelo menos um deles. (Expande).	buscar design | "ciências e matemática" mostrará as páginas que contém um dos termos ou ambos
NOT	!	O operador ! buscará registros que contêm o termo de pesquisa que o precede, mas não o termo que o sucede. (Restringe).	ensino de ciências ! ensino de matemática produzirá documentos relacionados a ao ensino de ciências, sem mostrar nenhum onde apareça também ensino de matemática.
NEAR	~	Encontra documentos contendo ambas as palavras especificadas.	
	( )	Os parênteses servem para elaborar pesquisas ainda mais complexas, definindo operações menores dentro da expressão inteira. A busca funciona nesse caso considerando os parênteses como se fossem termos isolados, e depois os combina.	"ensino" & (ciências matemática) & Brasil) . Apresentará páginas que contenham o termo ensino, o termo Brasil, ao mesmo tempo, pelo menos o termo ciências ou o termo matemática.

Fonte: PRIETO, Débora Durán Prieto. **Busca Booleana.** Universidade de São Paulo. Faculdade de Educação. Grupo de pesquisa: Projeto Telescola.

Disponível em: <http://www.conscienciologia.net/mecanismos_busca.htm>

Acesso em: 17 jul. 2005.

2.4.3 Truncagem

A truncagem ou busca por raiz é a possibilidade de busca por prefixo ou sufixo, para substituir uma letra ou conjunto de letras de uma palavra-chave. Ela pode ser feita através do uso do asterisco no lugar da letra ou palavra que desejamos encontrar. Esta não é uma característica comum dos mecanismos de busca. O Google, por exemplo, não utiliza a truncagem feita pelo uso do asterisco

no meio de uma palavra. Ele permite que se utilize, no entanto, o asterisco entre termos completos.

> Truncagem (*Truncation*) ou Coringa: É um meio de facilitar o usuário, muitas vezes utilizado pelos mecanismos na internet, lembrando o recurso utilizado nos Sistemas Operacionais tais como DOS, Windows, Unix, etc. Os símbolos adotados normalmente para o coringa são o asterisco (*), e, em alguns casos, o símbolo de interrogação (?). Se colocarmos o texto “car*”, para pesquisa teríamos como resultados possíveis: car, carro, carroça, Carlos, etc. (ANDRADE, 2004, p. 32)

A truncagem é muito útil quando desejamos encontrar uma palavra ou expressão sem saber sua ortografia correta. Para isso precisamos colocar o símbolo utilizado na truncagem no lugar da letra que estamos em dúvida.

2.4.4 Ordenação dos Resultados

Os critérios utilizados por cada mecanismo de busca fazem parte, na maioria das vezes, do sigilo da empresa que os mantém. Podemos notar diferenças entre os resultados quando executamos uma mesma busca em mais de um mecanismo de busca. Os resultados podem até ser os mesmos, ou seja, os *sites* recuperados podem ser os mesmos. No entanto iremos encontrá-los em posições diferentes na hora da apresentação dos resultados. A ordem ou posição dos *sites*, não será a mesma. Isso se deve à utilização de algoritmos de ordenação por parte dos mecanismos.

Entre os critérios mais utilizados por estes algoritmos, como coloca Céndon (2001, p. 45) estão “[...] a localização e freqüência de ocorrência das palavras em uma página”. Podem também ser levados em consideração o número de termos da consulta que estão presentes na página e a proximidade em que os termos se encontram. O tamanho do documento, os documentos curtos seriam mais

importantes que os documentos longos, esse critério é chamado de densidade. (CENDON, 2001)

2.5 Diretórios

Os diretórios foram os primeiros sistemas de busca da web inventados, eles foram a primeira solução encontrada para organizar e localizar os diversos recursos da *Web*. Porém, quando foi inventado, essa tarefa era muito mais fácil pois a quantidade de informações disponíveis na *web* era pequena, permitindo que fossem coletadas por seres humanos, de forma não automática. O primeiro diretório da *web* foi o *The World Web Virtual Library*, lançado em 1992 e sediado no CERN, mesmo local de nascimento da *web.* Atualmente, o *Yahoo!* é o diretório mais popular existente. Ele iniciou em 1994 a partir do trabalho de estudantes de doutorado da *Stanford University.*

Os diretórios organizam endereços eletrônicos por categorias e sub-categorias, ou seja, eles recebem uma **ordem hierárquica** de assunto, geralmente com termos bem amplos ou pouco específicos. Ex. educação, esporte, entretenimento, viagens, etc. (CENDON, 2001, p. 39) Ao navegar a procura de uma busca, devemos então, partir de assunto geral para assuntos cada vez mais específicos, até a resposta.

Podemos entender o que é um diretório e como ele funciona, imaginando uma biblioteca. O usuário vai até a biblioteca procurar um determinado assunto, quando ele chega lá, encontra os livros agrupados por assuntos. O usuário não quer pesquisar no catálogo dessa biblioteca, ele simplesmente quer ir até as estantes e, ao manusear os livros, encontrar o que está buscando, ou até algo mais. É semelhante à busca no diretório, porem, no diretório as estantes são representadas

pelas categorias e os livros representados pelos sites.

Os diretórios têm dois componentes principais, uma **base de dados**, também chamada de índice ou catálogo e um **programa de computador** que faz a pesquisa na base de dados. A montagem ou criação da base de dados de um diretório é realizada por humanos, são eles que fazem a análise e a indexação dos sites da web. Nos diretórios, não existem robôs para a catalogação e indexação das páginas da web, isso é feito pela equipe de editores que trabalham nas empresas de cada diretório.

Segundo MOURA (2001, p. 112) e MOTA (1998, p. 22) o sistema hierárquico de assunto, que funciona com perfeição nas unidades de informação, não é tão eficiente quando usado na internet, isso porque nas bibliotecas a classificação de assuntos é feita por profissionais bibliotecários que seguem padrões rígidos e mundialmente padronizados. Já na internet, a classificação é feita de acordo com as conveniências de cada diretório (ou empresa responsável pelo diretório) e não raramente a classificação é feita pelo autor da própria *home page*, o que demonstra uma falta de critérios padronizados de relevância. Alguns autores como Céndon (2001) e Aguilho (s.d.) não concordam com estas afirmativas e falam sobre como é feita a seleção de informações e quais os critérios de qualidade utilizados pelos diretórios para inclusão de um *site* em sua base de dados.

> O descobrimento e a seleção das informações é realizada em sua maioria por profissionais especializados, os editores (geralmente documentalistas, bibliotecários) que aplicam critérios de qualidade para avaliar se um site pode ser indexado ou não no diretório. Os editores descobrem novos sites a partir de sugestões do usuário (cadastro do site pelo usuário), através de pesquisas na Internet como listas de anúncios de novas páginas, ou ainda, pelo uso de robôs que coletam novas URLs na web. (CÉNDON, 2001, p. 3).

> Quanto aos critérios de qualidade utilizados para incluir um site em um diretório destacam-se os aspectos de legibilidade; a identificação (se existe correio eletrônico, se o nome do autor aparece na página dentre outros), a estruturação e a riqueza em multimídia. (AGUILHO, s. d., p. 3).

No caso dos diretórios, nem sempre a qualidade é o critério mais prezado. Alguns diretórios analisam primeiro os sites que lhes sejam submetidos junto com um pagamento. Outros apenas incluem sítios que paguem para neles serem incluídos. Também é freqüente que, na primeira página de resultados que apresentam, haja uma predominância de sites comerciais. O que compromete a busca da maioria das pessoas que analisa apenas as primeiras páginas de resultados.

2.5.1 Organização Hierárquica de Assunto

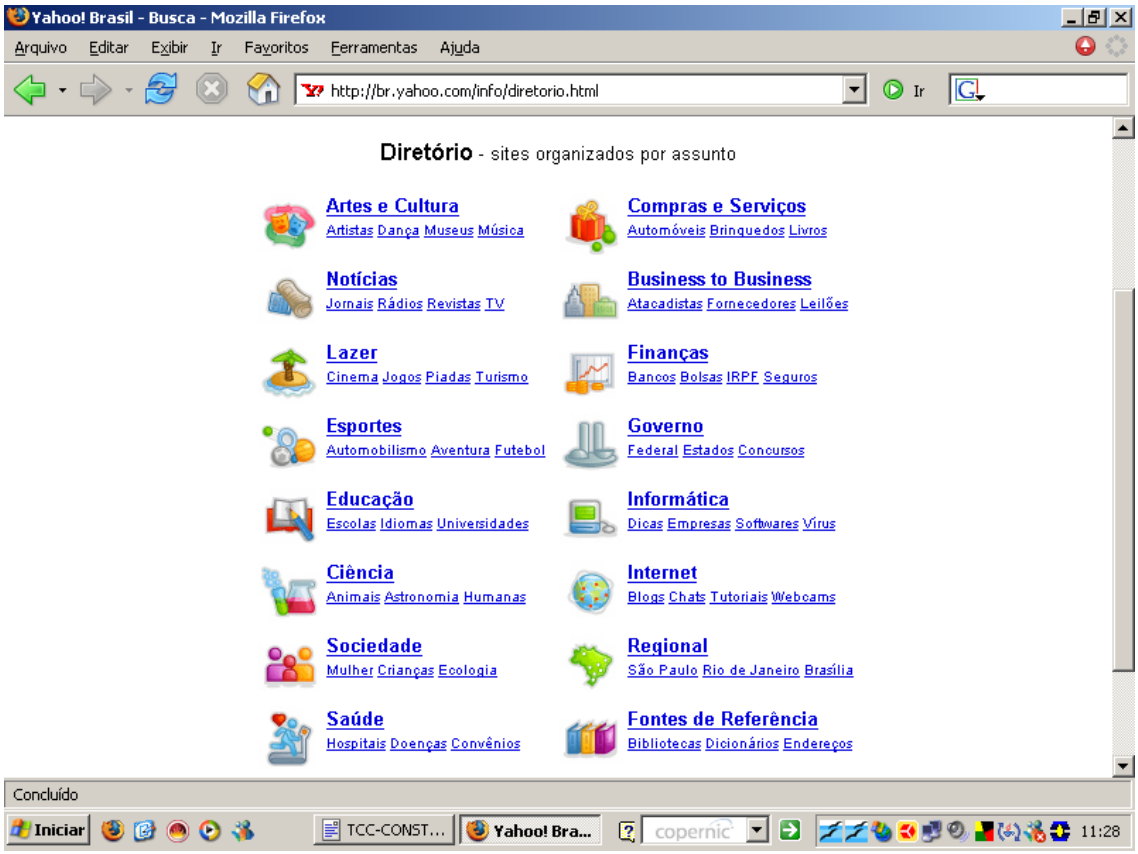
Em um sistema de classificação usado em unidades de informação como é o caso da CDU (Classificação Decimal Universal), em uma biblioteca universitária, por exemplo, o usuário encontrará os documentos organizados globalmente em: Generalidades; Filosofia; Religião, Teologia; Ciências Sociais; Lingüística; Ciências Matemáticas, Físicas e Naturais; Ciências Aplicadas, Medicina, Tecnologia; Belas-artes; Literatura; História e Geografia. Depois, estas categorias gerais são divididas em subcategorias obedecendo a uma ordem hierárquica de assunto. Se o usuário fizer a mesma pesquisa usando diretórios na internet, ele encontrará as informações organizadas também em grandes categorias e, depois, em sub-categorias, também obedecendo a uma ordem hierárquica de assunto. No entanto, os diretórios da web não utilizam notações numéricas como as encontradas na CDU. Também não são utilizadas exatamente as mesmas categorias globais, mas os princípios de

organização são os mesmos. No caso do *Yahoo!* e do *Cadê?*, por exemplo, as categorias são divididas em Artes e Cultura; Notícias; Lazer; Esportes; Educação; Ciência; Sociedade; Saúde; Compras e Serviços; *Business to Business*; Finanças; Governo; Informática; Internet; Regional e Fontes de Referência. (MOTA, 1998, p. 20). Existem também os diretórios temáticos ou especializados, alguns deles utilizam esquemas de classificação como a CDU, no entanto, estes diretórios são criados ou mantidos por profissionais da informação. (CÉNDON, 2001).

> A estrutura temática dos diretórios é essencial. Se a diversidade de temas e a sua organização forem intuitivas e lógicas, então é mais provável que o usuário encontre neles aquilo que procura. Se tal não suceder, até mesmo que o seu conteúdo disponha da resposta que o utilizador procura, é menos provável que essa resposta venha a ser encontrada. (ROSAS; HOURMANT 2002, p. 2).

Para ROSAS e HOURMANT (2002, p. 2), um bom diretório deveria poder substituir as enciclopédias, que também têm, freqüentemente, uma estrutura temática mas afirma que nenhum deles sequer se aproxima desse objetivo. Os diretórios fornecem respostas mais atuais do que as enciclopédias, além de dar mais ênfase a temas específicos como a informática e temas de interesse comercial. Geralmente as enciclopédias não abordam esses temas ou o fazem de forma superficial. As enciclopédias apresentam, geralmente, uma estrutura mais coerente e possuem temas mais clássicos e teóricos como informações sobre temas pertencentes a áreas específicas do conhecimento. A vantagem é somente os sistemas de busca permitem ao usuário, acesso imediato a bases de dados especializadas cujas informações, na maioria das vezes, somente interessam aos especialistas. Veja na figura abaixo, como a maioria dos diretórios esta organizada em sua hierarquia de assunto.

FIGURA 3 – Organização Hierárquica dos Diretórios



Disponível em: <http://br.yahoo.com/info/diretorio.html>. Acesso em: 21 de jul. 2005.

2.6 Metabuscadores

Os metabuscadores, também chamados de multibuscadores, permitem a busca em mais de um sistema de busca, diretórios ou motores. O metabuscador **traduz** a pesquisa do usuário nos termos usados por cada mecanismo e oferece a pesquisa pronta ao usuário, de forma unificada e padronizada. Existem também os pseudometabuscadores, mas eles apenas listam os buscadores, sem fazer a pesquisa para o usuário. Neste caso há uma caixa de pesquisa para cada buscador e as pesquisas são submetidas separadamente, e não simultaneamente como nos metabuscadores. Na maioria das vezes, os metabuscadores permitem com que o usuário escolha, dentre uma lista, quais os sistemas deverão ser utilizados e ao

mostrar os resultados, indica os sistemas de busca onde ele encontrou cada um dos *sites*. No caso de duplicações, o metabuscador se encarrega de eliminá-las, poupando o tempo do usuário.

Existem também, metabuscadores que podem ser instalados direto no computador do usuário, facilitando a construção local de estratégias de busca. São exemplos desta categoria o *Copernic* (http://www.copernic.com), e o *Web Ferret* (http://www.ferretsoft.com/netferret/_).

> O nível *meta search engines* proporciona a consulta simultânea a diversos sistemas de busca simples, ou seja, executa automaticamente buscas utilizando *simple search engines* de forma paralela fornecendo ao usuário o resultado das diversas buscas de forma resumida e simplificada em único formulário, tal como nas ferramentas mais simples. Este tipo de agente não possui banco de dados porque faz uso dos bancos de dados das diversas *simple search engines*. (ABREU, 2003, p. 18).

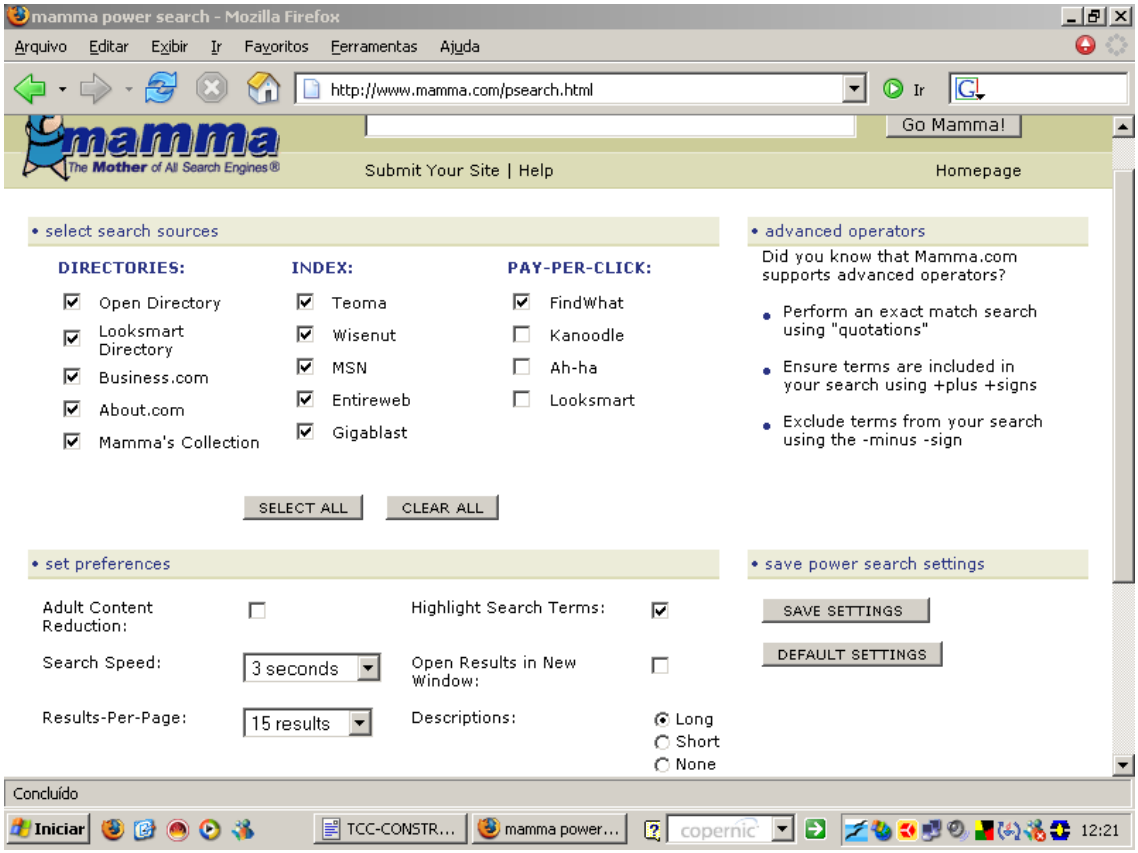
Apesar de parecer, a primeira vista, o sistema mais eficiente, os metabuscadores possuem algumas limitações/desvantagens que devem ser levadas em consideração. É o que nos mostra em seu trabalho sobre sistemas de busca a autora CÉNDON (2001, p. 49). “Os recursos de busca, específicos de cada motor, que são os mecanismos utilizados para um maior refinamento das pesquisas, tornam-se inacessíveis na interface do metamotor”. Além disso, devido ao grande volume de informações existentes na web, nos resultados obtidos, obtém-se maior quantidade de informações sem um correspondente aumento de qualidade. Em alguns metabuscadores apenas um subconjunto dos resultados de cada ferramenta é recuperado.

Devido a estas limitações os metabuscadores são mais indicados para buscas onde se utilizam termos únicos ou outras buscas simples, que não necessitem maior sofisticação. Além disso, o tempo de busca é sempre maior que o tempo utilizado nas buscas feitas através de outros sistemas, isso porque o

metabuscador irá realizar a busca diretamente nos diretórios e mecanismos, logo, a busca será tão demorada quanto o sistema mais demorado que o metabuscador utilizou. Ou ainda, o metabuscador irá estipular um tempo máximo que ele poderá esperar pela resposta do mecanismo. Ao ultrapassar este tempo, o metabuscador terminará a busca, às vezes ficando sem resposta do mecanismo que demorou.

Veja na figura abaixo, o exemplo do metabuscador "*Mamma*" (www.mamm.com) e como ele permite ao usuário escolher os sistemas de busca que serão usados para sua pesquisa:

FIGURA 4 – Metabuscador "Mamma"



Disponível em: <http://www.mamma.com/psearch.html>. Acesso em: 19 jul. 2005.

2.7 Sistemas de Busca: uma comparação

A tabela a seguir apresenta uma análise comparativa entre as principais características dos sistemas de busca e foi adaptada a partir da tabela mostrada no estudo de Deters e Adaime (2003).

TABELA 3 – Comparação entre Diretórios, mecanismos de busca e metabuscadores.

Sistema de Busca	Descobrimento das páginas	Representação do conteúdo do documento	Representação da consulta	Apresentação dos resultados	Tempo de Busca
Diretórios	Realizada manualmente (por pessoas)	Classificação manual	Implícita mediante navegação pelas categorias.	Página de resultados previamente construída. Os resultados são mostrados de forma bastante precisa.	Depende do conhecimento que o usuário possui a respeito do assunto procurado e da forma de organização do diretório.
Mecanismos de Busca	Principalmente de forma automática mediante *robots*	Indexação automática	Explícita mediante palavra-chave.	Página criada de forma dinâmica para cada consulta. Pouca precisão	Depende da rapidez do próprio mecanismo. Geralmente os mecanismos são os mais rápidos entre os sistemas de busca e fazem uma pesquisa, em média, a 0,05 segundos.
Meta-buscadores	Não possuem mecanismos de descobrimento próprio.	Usam a base de dados de outros sistemas de busca, não indexam o conteúdo.	Explícita mediante palavra-chave.	Páginas criadas de forma dinâmica apresentam uma maior cobertura, mas os resultados são pouco precisos.	O tempo de pesquisa do metabuscador é maior devido ao fato de ser necessário traduzir e compilar os resultados dos sistemas de busca onde ele faz as buscas.

A tabela a seguir, mostra algumas vantagens e algumas desvantagens de uso de cada sistema de busca, diretórios, mecanismos e metabuscadores. Foi desenvolvida a partir dos estudos feitos pelos autores Céndon (2001); Deters e Adaime (2003); Moura (2001); Dominguéz (2001) e Nahuz (1999).

TABELA 4 – Vantagens e Desvantagens dos Sistemas de Busca

SISTEMA DE BUSCA	VANTAGENS	DESVANTAGENS
DIRETÓRIOS	São mais fáceis de serem utilizados, principalmente para usuários leigos, pois é necessário apenas navegar pelas categorias.	Demora na atualização de informações, não há nenhum mecanismo automático que faça as suas atualizações. Atualizar manualmente as informações indexadas, torna-se uma tarefa impossível.
	Permitem ter uma visão geral do volume e conteúdo do índice, muitos diretórios indicam em cada um dos seus nodos quantas referências e subcategorias existem nela.	Possuem uma pequena cobertura da Web, ou seja, poucas páginas indexadas na sua base de dados.
	As informações disponíveis passaram por um processo de seleção de qualidade e com isso os resultados de uma pesquisa são mais precisos.	A seleção, a classificação e a descrição dos recursos na maioria dos casos são feitos por várias pessoas, o que conduz conseqüentemente a uma falta de critérios homogêneos.
	economizam tempo em buscas de informações irrelevantes	
MECANISMOS DE BUSCA	Possuem informações atualizadas diariamente.	Não possui nenhum tipo de vocabulário controlado.
	A cada ano as empresas que mantém os mecanismos investem muito dinheiro no aperfeiçoamento dos algoritmos.	Cada mecanismo tem a sua própria sintaxe para "expressar" a consulta, o que representa uma das grandes dificuldades para o usuário. Esses acabam utilizando sempre o mesmo mecanismo.
	Permitem pesquisas amplas.	Retornam resultados pouco precisos, sendo que, as informações indexadas não passaram por um processo de qualidade.
META-BUSCADOR	Poupa tempo do usuário quando realizada uma pesquisa simples. Ao invés de o usuário repetir a mesma pesquisa em vários sistemas ele pode utilizar apenas o metabuscador.	
	Realizam buscas em vários sistemas simultaneamente e acabam tendo uma cobertura bem maior da Web (nem todos os sistemas têm as mesmas páginas indexadas).	Por causa do grande volume de informações na internet, nos resultados obtidos, normalmente obtém-se maior quantidade de informações sem um correspondente aumento de qualidade.
	Possibilitam ao usuário escolher em quais sistemas de busca o sistema deverá efetuar a consulta.	Em alguns metabuscadores apenas um subconjunto dos resultados de cada ferramenta é recuperado;
	Existe a necessidade de aprender a usar uma única interface para realizar a consulta.	os recursos de busca específicos de cada motor, que são os filtros dos mecanismos, tornam-se inacessíveis. Devido a estas limitações os metabuscadores são mais indicados para buscas onde se utilizam termos únicos ou outras buscas simples, que não necessitem maior sofisticação.

3 METODOLOGIA

Este estudo tem caráter comparativo e está dividido em três momentos. Primeiramente foi feita uma revisão bibliográfica, com isto pretendeu-se saber mais sobre o funcionamento de cada um dos sistemas de busca. A segunda fase do trabalho se deu por uma comparação das ferramentas de busca[1], fazendo assim uma comparação entre os resultados obtidos, até que se chegasse a conclusão de quais sistemas eram mais eficientes na busca de recursos informacionais. Na terceira fase do trabalho os sistemas de busca foram experimentados quanto ao seu nível de precisão e revocação. Esta fase teve como objetivo, analisar a qualidade dos resultados obtidos nos diferentes sistemas.

3.1 Objetivo Geral

O objetivo geral deste estudo é identificar as características, diferenças e semelhanças entre os sistemas de busca, com o intuito de oferecer subsídios aos profissionais da informação para que os mesmos obtenham resultados mais eficientes nas suas pesquisas bibliográficas.

3.2 Objetivos Específicos

Foram definidos como objetivos específicos:

- conceitualizar e identificar características, diferenças e semelhanças, entre os sistemas de busca da internet, isto é, entre mecanismos de busca, diretórios e metabuscadores, com base na literatura;

1Neste estudo, o termo “ferramentas” é utilizado como sinônimo de “recursos” e esta relacionado a todos os sistemas de busca e não como sinônimo de mecanismos de busca.

- identificar e descrever os principais sistemas de busca de propósito geral disponibilizados na web;
- experimentar os instrumentos de pesquisa dos sistemas de busca identificados e comparar a relevância dos resultados obtidos.

3.3 Levantamento Bibliográfico

Foi realizado um levantamento bibliográfico, através de buscas feitas em diversos sistemas de busca na internet. Através dos documentos obtidos partiu-se para a busca, tanto na web como fora dela, dos materiais citados nas referências de cada um dos documentos obtidos. Este levantamento bibliográfico foi dividido nas seguintes fases:

a) Investigação exploratória: foi realizada baseada na leitura rápida do material bibliográfico, e teve por objetivo verificar a relevância de cada documento para o estudo. Esta investigação foi feita mediante o exame dos resumos e das referências. Também foram analisadas a introdução e conclusão. Com a análise destes elementos foi possível ter uma visão global da obra, bem como de sua utilidade para a pesquisa;

b) leitura seletiva: foi feita a seleção do material que de fato interessava à pesquisa. Como os objetivos geral e específico já haviam sido estabelecidos, foi relativamente fácil fazer a seleção das obras que realmente interessavam para o presente trabalho;

c) analise das referências e citações: foram analisadas as referências e citações dos materiais resultantes da leitura seletiva, e as julgadas relevantes foram então recuperadas e acessadas através da web ou fora dela. Com este procedimento desejava-se obter um maior número de resultados;

d) leitura analítica: por fim foi feita uma leitura analítica a partir dos textos selecionados. A finalidade desta leitura foi a de ordenar e sumariar as informações contidas nas fontes, de forma que estas possibilitassem a obtenção de respostas ao problema da pesquisa.

3.4 Identificação dos Sistemas de Busca

Para que o experimento entre os sistemas de busca fosse feito, foi necessário identificar dentre uma enorme quantidade de sistemas, aqueles que fossem, de certa forma, os mais importantes existentes na web. O critério seguido foi o de popularidade do sistema. Isso porque para que este trabalho seguisse parâmetros “reais” de busca, era necessária a utilização de sistemas que hipoteticamente estivessem sendo utilizados pela maioria dos profissionais da informação.

A identificação dos sistemas se deu pelo número de vezes que cada sistema foi citado em cada um dos textos consultados para construção deste trabalho. Dos mais citados, os quatro primeiros foram eleitos como os mais importantes e assim utilizados nas buscas existentes neste estudo. Dos diretórios, porém, foram identificados apenas 2 e não quatro como as demais categorias. Isso porque além destes dois o restante foi citado apenas uma vez cada um e por isso considerados irrelevantes.

Os autores utilizados para a contagem de citações foram:

- ✓ ABREU (2003);
- ✓ AIRES, ALUÍSIO (2003);
- ✓ ALENCAR (2001);
- ✓ ANDRADE (2004);
- ✓ BLATTMANN; FACHIN; RADOS (1999);
- ✓ BRANSK (2004);
- ✓ CÉNDON (2001);
- ✓ DETERS, ADAIME (2003);
- ✓ DOMINGUEZ (2001);
- ✓ GONZALES (2002);
- ✓ LOPES (2002);
- ✓ MARCONDES; GOMES (s. d.);
- ✓ MOTA (1998);
- ✓ MOURA (2001);
- ✓ NAHUZ (s. d.);
- ✓ NOVO; NOVO (1999);
- ✓ SANTOS (2001);
- ✓ SILVEIRA (2002);
- ✓ TEIXEIRA; SHIEL (1997);
- ✓ URBANO (s. d.).

3.5 Formulação dos Instrumentos para a Análise dos Sistemas de Busca

Na segunda fase do trabalho as ferramentas dos sistemas de busca foram experimentadas com o objetivo de se traçar comparações entre os resultados obtidos acerca da sua quantidade e qualidade. Esta fase foi dividida em:

a) investigação acerca da quantidade de resultados obtidos através de uma busca simples (apenas 1 termo). Teve como objetivo verificar o tamanho das bases de dados dos diferentes sistemas de busca;

b) a busca realizada através de expressões foi verificada, com ou sem o uso de aspas, verificando a quantidade e qualidade dos resultados obtidos com o intuito de entender melhor a linguagem de busca adotada por cada sistema;

c) o uso da truncagem foi verificado, utilizando um termo ou uma expressão e teve como objetivo observar como cada sistema se comporta com o uso de caracteres coringa e quais os símbolos cada um adota;

d) foi feita uma investigação acerca dos resultados obtidos com palavras homógrafas com o objetivo de criar estratégias de busca para a especificação do assunto procurado;

f) finalmente, os sistemas de busca foram experimentados quanto ao uso ou não de linguagem natural. Essa investigação teve como objetivo verificar se os sistemas permitem ao usuário fazer perguntas diretas, descrevendo as informações que desejam encontrar.

3.6 Formulação dos Instrumentos para Análise da Precisão e Revocação dos Sistemas de Busca

Na terceira fase do trabalho os sistemas de busca foram experimentados quanto ao seu nível de precisão e revocação. Esta fase teve como objetivo, analisar a qualidade dos resultados obtidos nos diferentes sistemas. Foi baseada na metodologia adotada pelos autores SHAFI e RATHER (2005), presente no artigo “Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology” e foi dividida da seguinte forma:

a) Dos 10 sistemas de busca apresentados nestes trabalhos, 6 foram selecionados para este experimento. O critério seguido na escolha desses sistemas foi o mesmo adotado no item 4.1 (Identificação dos Sistemas de Busca), ou seja, foram escolhidos os 2 sistemas mais citados pertencentes a cada categoria (mecanismo, diretório ou metabuscador).

b) O termo de busca foi escolhido por tratar-se de um termo utilizado no idioma inglês. Não prejudicando assim, os sistemas que possuem em suas bases de dados, uma predominância de documentos no idioma inglês.

c) Depois de eleito o termo de busca, as buscas foram feitas e o material recuperado no resultado foi coletado para análise.

d) Depois da análise das cópias dos resultados, alguns sistemas foram revisitados. Isso ocorreu sempre que houve problemas relacionados ao acesso aos links.

e) Para a determinação da relevância dos documentos recuperados não foi atribuído nenhum tipo de pontuação. Apenas sim para os relevantes e/ou não para os não relevantes. As páginas consideradas relevantes tratavam apenas do assunto pesquisado ou continham artigos científicos ou links sobre o assunto. Os documentos considerados não relevantes ou irrelevantes eram compostos por páginas pessoais, produtos comerciais, links repetidos ou apenas citavam o assunto pesquisado.

f) Foi confeccionada uma tabela com os dados obtidos na análise dos 6 sistemas para uma melhor visualização dos resultados.

g) Foi calculado o valor de precisão e revocação de cada sistema utilizado, com o objetivo de investigar sobre a qualidade da informação recuperada. De todos os resultados obtidos, foram analisados os 15 primeiros de cada sistema.

4 APRESENTAÇÃO E ANÁLISE DOS DADOS

4.1 Identificação dos Sistemas de Busca

Os sistemas de busca, identificados como os mais “importantes” ou mais populares, foram, como consta na metodologia, escolhidos com base no número de vezes que foram citados na bibliografia utilizada neste trabalho, são eles:

TABELA 5 – Identificação dos Sistemas de Busca

Mecanismo	Nº citações	Endereço Eletrônico
Altavista	20	Http://www.altavista.com/
Google	17	Http://www.google.com.br/
Excite	16	Http://www.excite.com/
Lycos	12	Http://www.lycos.com/
Diretório	Nº citações	Endereço Eletrônico
Yahoo!	18	Http://br.search.yahoo.com/dir
Open Directory Project	6	Http://dmoz.org
Metabuscador	Nº citações	Endereço Eletrônico
Metacrawler	8	Http://www.metacrawler.com/
Mamma	4	Http://www.mamma.com/
Dogpile	4	Http://www.dogpile.com
Copernic	3	Http://www.co pernic.com

Veja algumas características e um pouco da história de cada um dos sistemas:

4.1.1 Altavista

Foi desenvolvido em 1995, no Laboratório de Pesquisa Digital, Palo Alto, Califórnia, por uma equipe de técnicos e colocado em uso a partir de 15 de dezembro do mesmo ano. Permite acessar mais de 30 milhões de páginas Web em mais de 275 mil servidores e três milhões de artigos de 14 mil Usenet news groups.

É uma ferramenta poderosa, conhecida e usada mundialmente por milhões de pessoas, considerada por muitos, a "memória de elefante da Internet " pela sua abrangência. Utiliza robôs para manutenção dos sites, oferece pesquisa simples, por palavra-chave e avançada por frase ou expressão. Elimina pontuação e caracteres especiais como, $, %, /, _, &, etc. O Altavista, além de ser um mecanismo de busca, possui também um diretório, organizado numa ordem hierárquica de assunto, no entanto, este trabalho o analisa apenas como mecanismo de busca, desconsiderando seu diretório.

4.1.2 Google

O termo Google surgiu da palavra Google, inventada pelo Dr. Edward Rasner, da Universidade de Columbia para designar o número representado pela centésima potência do número 10, ou um número 1 seguido de 100 zeros. O termo Google pretende refletir a missão da empresa de organizar o enorme número de informações disponíveis na Web e no mundo. O Google é essencialmente um mecanismo de busca de palavras e links, ele recebe em média 200 milhões de consultas por dia e utiliza diversos recursos de filtragem e catalogação de resultados. O que garantiu o seu sucesso, já que existia inúmeros buscadores, foi, além de seus algoritmos de extração de dados, que tornam qualquer busca significativamente mais rápida do que qualquer procura realizada com outros sistemas de busca, a interface. O front-end do google faz com que a página retorne buscas quase imediatamente. (SEGREDOS, 2004). Além do mecanismo propriamente dito, o Google possui:

4.1.2.1 Google Diretórios

O Google diretórios (http://www.google.com.br/dirhp?hl=pt-BR&tab=gd&q) é muito parecido com o Yahoo diretórios, ele organiza o conteúdo da internet em uma ordem hierárquica de assunto.

4.1.2.3 Google News

O Google News (http://news.google.com) realiza buscas em serviços de notícias e, infelizmente, apenas no idioma inglês. Ao mostrar os resultados, há uma indicação de horário ao lado do link da notícia que mostra o horário em que a notícia foi colocada na web, ou seja, o quanto atual ela é.

4.1.2.4 Google Linux

O Google Linux (http://www.gooogle.com/linux) é um diretório voltado apenas para a busca de assuntos relacionados ao universo linux. É encontrado apenas no idioma inglês, apesar de recuperar resultados em outros idiomas, inclusive em português.

4.1.2.5 Google Microsoft

O Google Microsoft (http://www.google.com/microsoft.html) realiza buscas em todas as páginas relacionadas à empresa de Bill Gates. Neste diretório podemos encontrar não só dicas, mas também drivers para dispositivos específicos.

4.1.3 Excite

Criado recentemente pela Architext Software, oferece duas ferramentas de navegação em separado: NetSearch, que permite procurar, simultaneamente, os grupos de notícias Usenet via texto simples ou pelo teclado e o endereço Web; NetReviews, que oferece artigos Web organizados por assuntos.

4.1.4 Lycos

É um dos mecanismos de busca mais antigos da internet. Foi fundado em junho de 1995. Em outubro de 2000, nasceu a Terra Lycos, a partir da união da Terra Network S/A com a Lycos Inc., e em 2004 o Terra vendeu o Lycos. O Lycos possui versões em diversos idiomas. Faz buscas na web, busca por imagens,

notícias, vídeos, mp3, páginas amarelas e busca avançada. Possui também um diretório, no entanto, neste trabalho será considerado somente seu mecanismo de busca. Uma curiosidade do Lycos em espanhol, é o “Ahora están buscando”. É um link onde o usuário pode ver as palavras-chaves das buscas que estão sendo feitas no buscador, em tempo real. É só clicar em um termo da lista e os resultados são rapidamente disponibilizados.

4.1.5 Yahoo!

O Diretório Yahoo foi criado em abril de 1994 por David Filo e Jerry Yang. Ambos eram estudantes de engenharia elétrica da Universidade de Stanford. O trabalho começou com a catalogação de servidores Web, e tinha o objetivo de ajudar as pessoas a navegarem sobre as informações existentes na rede. A invensão tornou-se popular rapidamente pois oferecia um mecanismo de busca associado a uma lista de sites organizados por assunto. Além de ser um diretório, o Yahoo! possui também um mecanismo de busca, no entanto, este trabalho o analisa apenas como diretório, desconsiderando seu mecanismo de busca.

4.1.6 Open Directory Project (DMOZ)

O Open Directory Projet (Projeto de Diretório Aberto), conhecido também por ODP ou Dmoz (Directory Mozilla) é um projeto que conta com a colaboração de vários voluntários que editam e categorizam páginas da internet. Qualquer pessoa pode sugerir páginas que deverão ser aprovadas ou não pelo editor equivalente da categoria de envio. Qualquer pessoa pode também se candidatar a editor do dmoz através do preenchimento de um formulário com suas informações pessoais. Este formulário é aprovado ou não por um editor meta. (WIKIPÉDIA, 2005) No site <*http://dmoz.org/World/Portugu%C3%AAs/add.html*> é possível conhecer as etapas

existentes até a aprovação de inserção de site e suas normas editorias.

4.1.7 Metacrawler

Foi desenvolvido em 1994 na universidade de Washington pelo estudante Erik Selberg e pelo professor Oren Etzioni. Em 2000 se juntou com a InfoSpace. Faz buscas nos sistemas: Google, Yahoo!, Ask Jeeves, About, LookSmart, Overture e FindWhat.

4.1.8 Mamma

Foi desenvolvido em julho de 1996, por Herman Tumurcuoglu, na Universidade de Carleton de Ottawa. Pesquisa em diretórios e mecanismos de busca, são eles: Open Directory, Business.com, About.com, Google, Yahoo, Gigablast, Wisenut, MSN, Entireweb e Looksmart. Faz buscas na Web por notícias, imagens e páginas amarelas. Seu algorítmo é denominado “rSport”. Ao invés dele eliminar os resultados duplicados dos motores de busca e diretórios, ele considera cada resultado duplicado da busca e atribui uma “pontuação” para esse resultado. As páginas mais pontuadas são disponibilizadas no topo da lista de resultados.

4.1.9 Dogpile

Utiliza os sistemas: Yahoo, Lycos, Excite, WebCrawler, InfoSeek, AltaVista, HotBot, Google, MSN e LookSmart. Possui buscas na web por imagens, áudio, vídeos, notícias e páginas amarelas (ou comerciais). Possui filtros de busca como idioma, data e domínio.

4.1.10 Copernic

O Copernic é um pouco diferente dos outros metabuscadores, isso porque é possível instala-lo diretamente em nosso computador. Ele é distribuído em duas versões, uma gratuita que realiza buscas na Web, Usenet News e catálogos de pessoas e outras duas versões pagas, que oferecem a possibilidade de buscas em vários outros tipos de serviços de informação. O Copernic permite que as buscas sejam salvas e reexecutadas quando necessário. O Copernic remove entradas duplicadas e pode memorizar a busca e atualizar os seus resultados periodicamente, uma função que os metamotores de pesquisa normais não possuem. Ele proporciona buscas na web, busca de pessoa física (através de nome completo, cidade e estado e busca de pessoa jurídica (através do tipo de atividade, cidade e estado)).

Segundo Rosas e Hourmant (2001, p. 5) nas experiências de busca que fizeram utilizando o Copernic 2001 (versão gratuita):

> [...] este programa não foi superior ao Google em número de respostas, nem sequer com interrogações relativamente simples. Além disso, não dispõe da enorme variedade de filtros que o Altavista ou o Google proporcionam. Dispõe, porém, do operador NEAR, que o Google não suporta, embora nos nossos testes ele nem sempre tenha funcionado devidamente. O processo de eliminação de ligações quebradas pode demorar um tempo significativo. Em nosso entender, as suas respostas foram geralmente menos pertinentes do que as que obtivemos com o Google, embora mais pertinentes do que as de outros motores. Este programa pode, porém, substituir com vantagem os metamotores que conhecemos. Além disso, suporta funcionalidades neles não existentes: múltiplos tipos de pesquisas temáticas; resumo de páginas encontradas, para o que é necessário comprar um módulo adicional; tradução de páginas, através de um serviço de tradução em linha, cuja qualidade é muito inferior à de uma tradução humana; e memorização das pesquisas efetuadas, que pode repetir, nos casos em que consideremos importante atualizar os dados já obtidos.

4.2 Características dos Sistemas de Busca

As tabelas apresentadas a seguir foram confeccionadas com base no modo de apresentação dos resultados dos sistemas de busca utilizados. Nas ferramentas de "busca avançada" disponíveis por cada sistema e em algumas características consideradas relevantes, apresentadas no referencial teórico deste estudo.

TABELA 6 – Características dos Mecanismos de Busca

	Altavista	**Google**	**Excite**	**Lycos**
Categoria	Mecanismo	Mecanismo	Mecanismo	Mecanismo
***Abrangência (páginas)**	350 milhões	560 milhões	250 milhões	50 milhões
Indexa texto completo?	Sim	Sim	Sim	Sim
Busca em línguas específicas?	35 idiomas	35 idiomas	10 idiomas	10 idiomas
Procura frase exata?	Sim	Sim	Sim	Sim
Possui versão em mais de 1 idioma?	26 idiomas	116 idiomas	7 idiomas	Apenas inglês
Faz pesquisa avançada?	Sim	Sim	Sim	Sim
Faz pesquisa avançada *boolena*?	Sim	Sim	Sim	Sim
Procura por nome de domínio?	Sim	Sim	Sim	Sim
Procura imagens?	Sim	Sim	Sim	Sim
Procura *links*?	Sim	Sim	Não	Não
Procura por data?	Sim	Sim	Sim	Sim
Diferencia maiúscula e minúscula	Sim	Não	Não	Não
É possível pedir inclusão de *site*?	Sim	Sim	Sim	Sim
O termo de busca esta destacado no resultado?	Sim	Sim	Nem sempre	Nem sempre

* Dados retirados da tabela 1 (Tamanho da base de dados dos motores de busca).

TABELA 7 – Características dos Metabuscadores

	Metacrawler	**Mamma**	**Dogpile**	**Copernic**
Categoria	Metabuscador	Metabuscador	Metabuscador	Metabuscador
Sistemas que utiliza (diretórios e/ou mecanismos)	Google, Yahoo! Search, MSN Search, Ask Jeeves, About, MIVA, LookSmart.	Open Directory, Looksmart, Ah-ha.	Google, yahoo, Msn e Ask.	Não menciona
Que tipo de buscas faz (além da web)?	Imagens, áudio, vídeo, notícias e sites comerciais.	Novidades, imagens e sites comerciais.	Imagens, áudio, vídeo, novidades e sites comerciais.	Imagens, áudio, vídeo, novidades e sites comerciais.
Busca em línguas específicas?	sim	não	sim	sim
Procura frase exata?	sim	não	sim	sim
Possui versão em mais de 1 idioma?	não	não	não	não
Faz pesquisa avançada?	sim	sim	sim	sim
Faz pesquisa avançada *boolena*?	sim	não	sim	não
Procura por nome de domínio?	sim	não	sim	sim
Procura *links*?	não	não	não	sim
Procura por data?	sim	não	sim	sim
O termo de busca esta destacado no resultado?	Nem sempre	não	Nem sempre	Nem sempre
Mostra onde os resultados foram encontrados?	sim	sim	Sim	sim

TABELA 8 – Características dos Diretórios

	Yahoo	**Dmoz**
Categoria	Diretório	Diretório
Possui diretório e busca por palavras chaves?	Sim	Sim
Possui versão regional	Sim	Sim
Possui versão em mais de 1 idioma?	Sim	Sim

4.3 Avaliação dos Resultados de Busca

Esta fase do trabalho pretende investigar através da experimentação, diversas ferramentas existentes nos sistemas de busca. As experimentações foram feitas no período de agosto a outubro de 2005.

4.3.1 Investigação Acerca da Quantidade de Resultados

Esta investigação teve como principal objetivo, analisar a quantidade dos resultados obtidos através de uma consulta simples. A palavra chave utilizada para a consulta foi "Internet". A escolha se deu, primeiramente, por ser um termo representado igualmente, tanto no idioma português quanto no idioma inglês. Além desta característica, foi escolhido por ser um termo genérico e bastante difundido. Os resultados estão representados em ordem decrescente de resultados como mostra a tabela 9.

TABELA 9 – Quantidade de resultados obtidos

Sistema	*Nº de Resultados obtidos*
Altavista	**2.910.000.000**
Google	**2.480.000.000**
Lycos	**398.290.006**
DMOZ	**42.396**
Yahoo!	**7.692**
Dogpile	**120**
Metacrawler	**119**
Mamma	**94**
Copernic	**74**
Excite	**72**

Legenda:
[ ] Diretórios [ ] Mecanismos de Busca [ ] Metabuscadores

Nesta tabela podemos verificar que os mecanismos de busca encontraram um maior número de resultados, com exceção do Excite que foi o último colocado, em relação a todos os outros sistemas. Como já foi falado anteriormente e por isso já esperado, os mecanismos de busca possuem as maiores bases de dados dentre os sistemas de busca. Comprovamos isto através deste experimento. No entanto, os metabuscadores, (que tem por objetivo, encontrar um maior número de resultados, já que realiza a pesquisa nas bases de dados de vários mecanismos simultaneamente) encontraram menos resultados até que os diretórios (conhecidos por possuírem uma base de dados pequena). Isso confirma o que dizem os autores Rosas e Hourmant (2001, p. 5):

> Os metamotores nem sempre dão mais respostas do que os motores de pesquisa simples. Assiste-se até ao paradoxo, [...] de poderem dar menos respostas do que as obtidas com a utilização exclusiva de um só motor de pesquisa simples, mesmo quando utilizam tal motor para procurar responder à nossa interrogação.

Os mesmos autores (ROSAS; HOURMANT, 2001, p. 5) explicam que este fato pode ocorrer por três motivos:

a) os metabuscadores podem traduzir mal a interrogação para a sintaxe utilizada pelos mecanismos de busca ou diretórios;

b) os metabuscadores impõem limites de tempo às respostas que pesquisam nos outros sistemas de busca, por isso às vezes podem aproveitar apenas uma parte do conteúdo de cada um deles;

c) pode acontecer que nem sequer obtenham uma resposta de alguns dos motores que teoricamente consultam, caso em que por vezes, nos indicam que tais motores estão "timed out".

4.3.2 Investigação da Busca feita através de Expressões

Nem sempre, os termos utilizados em uma pesquisa através de sistemas de busca são feitos por meio de palavras simples. Na maioria das vezes são expressões que representam algum conceito dentro de uma determinada área do conhecimento. Este item procura investigar a maneira que os diferentes sistemas de busca interpretam a utilização de expressões. A maneira mais utilizada, tratando-se de sistemas de busca, é a utilização da expressão entre aspas, significando que os termos não podem ser recuperados separadamente ou em uma ordem diferente do que a representada pelo usuário.

> Esta função considera a hipótese de que quanto mais perto dois termos estejam dentro de um único texto, maior a probabilidade de estarem relacionados ao mesmo conceito. Nos mecanismos de busca na Web não é comum o uso do operador NEAR, no entanto, a busca restrita para uma expressão, quando disponível, costuma ser feita através do uso de aspas. (ALENCAR, 2001, p. 42).

Vejamos o exemplo de busca por expressões através do uso de aspas. A expressão utilizada na pesquisa foi “History of the Internet”. Veja na tabela abaixo os resultados obtidos nos diversos sistemas de busca.

TABELA 10 – Busca por Expressões

Sistema	***Nº de resultados obtidos sem o uso de aspas***	***Nº de Resultados obtidos com o uso de aspas***
Altavista	**293.000.000**	**786.000**
Google	**339.000.000**	**439.000**
Lycos	**64.330.005**	**64.390.005**
Excite	**82**	**80**
Dogpile	**85**	**66**
Metacrawler	**80**	**69**
Mamma	**83**	**66**
Copernic	**79**	**64**

Com base nos resultados acima, é extremamente importante na maioria dos casos, o uso de aspas como filtro de informação. Além de uma considerável diminuição na quantidade dos resultados, houve também um proporcional aumento na qualidade dos resultados. Isso porque através das aspas, os sistemas não recuperaram resultados onde as palavras da busca estivessem separadas, alterando o sentido da expressão. Com exceção do Excite, que, ao que tudo indica, não possui esta ferramenta. Ele obteve praticamente os mesmos resultados tanto com o uso de aspas como sem o uso. Além disso na busca com aspas o resultado não estava de acordo com o esperado. A expressão não estava agrupada no resultado e por isso os resultados foram insatisfatórios.

Os diretórios não foram investigados quanto ao uso de aspas, isso deve-se ao fato de não ser possível realizar buscas através de palavras-chaves. No entanto, a busca por **History of the Internet** foi realizada com o intuito de verificar a quantidade e qualidade das informações. Veja os resultados na tabela 11.

TABELA 11 - Diretórios: quantidade e qualidade de resultados

Diretório	**Links**	**Links iguais**	**Resultados Irrelevantes**	**Resultados Relevantes**
DMOZ	110	9	História de provedores (surgimento).	Documentos acerca da história dos computadores, da internet, da web, do CERN e dos sistemas de busca. História da internet em ordem cronológica, linha do tempo, conceitos e links interessantes, além de projetos ligados a história da internet.
Yahoo!	27	9	História de programas de computador e de pessoas ligadas a internet.	Documentos acerca da história, evolução e impacto da internet e da world wide web. Possuíam linha de tempo e alguns remetiam para outros sites que possuíam uma lista de links sobre a história da internet.

4.3.3 Investigação acerca do uso de truncagem

O experimento se deu, com o uso do termo “educa?” ou “educa*” ou ainda “educa$”. Os resultados advindos da truncagem poderiam ser em torno de: educa, educar, educação, educacional, educado(a), educativo(a), educador(a), educando(a). Dos mecanismos de busca, experimentados acerca do uso de truncagem, veja os resultados na tabela 12:

TABELA 12 – Resultados da Truncagem Simples

Palavra-chave / Sistema	Educa* ou educa?	educação	educacional	educado	educativo	educador	Educando
Altavista	4.450.000	21.900.000	3.500.000	1.140.000	10.600.000	1.280.000	655.000
Google	7.040.000	10.300.000	2.550.000	1.160.000	6.790.000	2.360.000	742.000
Excite	87	21	84	73	79	69	20
Lycos	617	1.485.000	482.200	171.200	4.946.000	285.900	108.300
Metacrawler	90	71	102	80	89	78	74
Mamma	84	62	77	60	76	73	66
Dogpile	95	73	93	83	84	76	78
Copernic	88	67	88	73	78	69	70

Podemos verificar através dos resultados, que embora alguns sistemas descrevam esta ferramenta no serviço de ajuda, nenhum deles mostrou resultados que comprovem isso. Os resultados provenientes das palavras truncadas (educa* ou educa?) deveriam ser aproximadamente a soma de todos os outros resultados (educação, educacional, educado, educativo, educador e educando) o que não acontece. Ao contrário, na maioria dos sistemas os resultados truncados foram menores que os resultados das outras palavras. Portanto, não podemos considerar o uso desta ferramenta, da forma descrita, em nenhum dos mecanismos utilizados neste estudo. No entanto, podemos utilizar a truncagem, no meio de expressões,

substituindo palavras inteiras e não apenas algumas letras. Veja os resultados da investigação do uso de truncagem em expressões na tabela.

TABELA 13 – Resultados da Truncagem por Expressão

Palavra-chave Mecanismo	Construção * no Brasil	Exemplos de palavras recuperadas no lugar do "*" (apenas os 10 primeiros resultados)
Altavista	4.140.000	Sociológica, naval, civil, cidadania, índia, etc.
Google	1.970.000	Refinaria, conhecimento, civil, direito, naval, etc.
Excite	22	Refinaria, cidadania, democracia, naval, civil, espacial, etc.
Lycos	339.400	Naval, presidiária, espaço, classes perigosas, igualdade, etc.
Metacrawler	78	Civil, refinaria, cidadania, naval, presídios, da nação, etc.
Mamma	35	Civil, naval, siderurgia, etc.
Dogpile	78	Civil, refinaria, cidadania, naval, nação, classes perigosas, democracia, etc.
Copernic	69	Cidadania, refinaria, civil, naval, democracia, etc.

Em todos os sistemas investigados, podemos perceber que há a possibilidade de truncagem feita entre uma expressão. Todos os mecanismos obtiveram resultados positivos acerca do objetivo da pesquisa. Podemos notar, entretanto, a diferença das palavras-chaves encontradas em cada um dos mecanismos. O que demonstra que cada um possui um fator de relevância diferente do outro o que faz com que os documentos recuperados estejam numa ordem diferente de importância. Podemos lembrar também que os mecanismos podem não indexar as mesmas páginas ou sites, o que poderia gerar este resultado. No caso dos metabuscadores, podemos verificar, mais uma vez, que a quantidade de resultados é inferior em relação aos mecanismos de busca e as informações obtidas são praticamente as mesmas.

4.3.4 Investigação acerca das buscas realizadas com palavras Homógrafas

Pesquisas por palavras-chave buscam *sites* em que apareçam as palavras consideradas importantes. O problema é quando a palavra tem mais de um significado, ou seja, são homógrafas. É o caso, por exemplo, da palavra "servidor", que pode significar funcionários públicos ou servidores web. Neste caso, é necessário associar palavras para, como diz Moura (2001), aumentar o índice de relevância da busca.

> [...] aumentar o índice de relevância é usar duas ou mais palavras, uma frase [...] com isso, aumenta-se a compreensão do conceito pesquisado, trazendo, por conseguinte, uma redução da extensão do universo a que o conceito se aplica. (MOURA, 2001, p. 6)

Por exemplo, quando procurarmos por servidores públicos, devemos digitar: “**servidor publico**” “**governo federal**” **Brasil**, isso faz com que aumentemos a relevância de nossa busca. Veja a diferença nos resultados apresentados na tabela 14:

TABELA 14 – Palavras Homógrafas

	Palavra-chave	***Palavra-chave***
Sistema	***Servidor***	“**servidor publico**” “**governo federal**” **Brasil**
Altavista	16.500.000	68.600
Google	8.990.000	128.000
Lycos	4.631.000	99
Excite	30	70
Dogpile	85	29
Metacrawler	85	80
Mamma	92	45
Copernic	76	13

O Copernic coloca automaticamente aspas na expressão como um todo. Ele executa a pesquisa da seguinte forma: ““servidor publico” “governo federal” brasil”. Para que o resultado fosse satisfatório, no entanto, a pesquisa foi realizada sem as aspas em servidor, o que resultou em “servidor publico” “governo federal” brasil”. Então pode-se notar a diferença entre os resultados, que diminuíram muito com a pesquisa por mais de um termo.

O Excite também demonstra inserir aspas automaticamente nos termos de busca. O mesmo procedimento feito no Copernic foi adotado. No entanto a quantidade de resultados aumentou de 30 para 70. O que pode demonstrar uma falha na configuração das ferramentas do mecanismo. Mesmo com aspas (inseridas pelo usuário ou automaticamente pelo sistema) o mecanismo não recupera as expressões juntas. Ele recupera apenas os termos individualmente. Podemos dizer então que ele não possui opção de uso de aspas.

O metacrawler também coloca aspas automaticamente e mesmo assim obteve uma pequena diferença na quantidade de resultados. O metabuscador recuperou os termos de pesquisa de modo individual e não os interpretou como uma expressão. Este fato pode ser explicado por haver alguma falha na tradução da pesquisa feita para o entendimento dos buscadores utilizados pelo metacrawler. Nos outros sistemas de busca, os resultados foram satisfatórios, uma vez que a quantidade de resultados diminuiu e a qualidade aumentou.

4.3.5 Investigação do uso de Linguagem Natural (LN)

Outra opção para a recuperação da informação e a relação entre os termos é o uso da linguagem natural, característica que permite ao usuário fazer uma pergunta direta, descrevendo a informação que deseja encontrar. Essa é uma possibilidade que alguns mecanismos de busca começam a utilizar, sendo apontada por alguns autores como uma das tendências para facilitar o trabalho dos usuários. (ALENCAR, 2001). Os tipos de perguntas mais utilizadas neste tipo de busca são: onde, o que é, em quais, o que é e para que, qual, como, aonde, quais, por que e quanto. (AIRES, 2003). No entanto, segundo Cardoso e Oliveira (s.d., p. 9), o uso de linguagem natural é ainda pouco utilizado em Mecanismos de Busca, eles dizem que muitos sistemas até possibilitam ao usuário escrever em LN (Linguagem Natural) a frase de consulta, mas utilizam stopword (sâo palavras que nâo devem fazer parte do termo na pesquisa, sâo irrelevantes) para proceder a busca.

Veja na tabela 15 os resultados das buscas feitas em linguagem natural, utilizando a pergunta “**o que é stopword?**”.

TABELA 15 – Linguagem Natural (LN)

Sistema	***Nº de resultados do uso de LN***	***Relevância dos resultados*** ***(10 primeiros resultados)***
Altavista	102	Apenas 4 resultados responderam a questão. Entre os outros resultados, 4 eram duplicados e não respondiam a pergunta, 1 era um link inativo e 1 apenas citava o termo sem dar seu conceito/significado.
Google	529	4 resultados responderam a questão, o primeiro fazia parte de um glossário na área da informática. 4 eram arquivos mortos. 2 apenas citavam o termo.
Lycos	21	4 resultados responderam a questão, o primeiro também fazia parte de um glossário na área da informática. 1 era um arquivo morto. 5 apenas citavam o termo sem defini-lo.
Excite	27	8 resultados responderam a questão, o primeiro também fazia parte de um glossário na área da informática. 2 apenas citavam o termo e eram duplicados.
Dogpile	32	6 resultados responderam a questão, o primeiro também fazia parte de um glossário na área da informática. 3 apenas citavam o termo.
Metacrawler	24	7 resultados responderam a questão. 2 apenas citavam o termo. 1 nem se quer cita o termo.
Mamma	23	4 resultados responderam a questão. 5 apenas citavam o termo. 1 nem sequer citava o termo.
Copernic	19	9 resultados responderam a questão. 1 apenas citava o termo de busca.

Dentre os resultados explicitados acima, destaca-se o Copernic e o Excite no uso de linguagem natural. Ambos obtiveram uma pequena quantidade de resultados, no entanto, a maior parte desses resultados pôde ser considerada relevante. O Altavista, Google e Lycos, obtiveram apenas 4 resultados relevantes. O altavista e o Google apesar de terem recuperado uma maior quantidade de resultados, tiveram a maioria considerada irrelevante.

4.3.6 Precisão e Revocação

Precisão e revocação são termos bastante utilizados na área da informática. São conhecidos como Precision e Recall, poucos autores utilizam suas traduções para o português. “A precisão é a porcentagem dos itens recuperados que são relevantes. A revocação é a porcentagem dos itens relevantes que foi recuperada” (SANTOS; NASCIMENTO, s.d., p. 115), ou ainda, como diz Souza (2004, p. 3) revocação é a “Razão do número de documentos atinentes recuperados sobre o total de documentos atinentes disponíveis na base de dados” e precisão a “Razão do número de documentos atinentes recuperados sobre o total de documentos recuperados.

Na web, existe quase que uma baixa qualidade da indexação generalizada quando a busca é feita através de mecanismos e metabuscadores. Isso porque ela é feita automaticamente. Isso faz com que recuperemos uma grande quantidade de informações a cada busca. Muitas destas informações, não possuem nenhuma relevância. Já nos diretórios, a indexação é feita através do trabalho humano e seguindo, na maioria das vezes, padrões de indexação. Podemos dizer então que em termos de recuperação de informação, os diretórios apresentam, teoricamente, maior precisão e relevância e os motores de busca maior revocação (SANTOS; SILVA, s.d.).

Existem dois fatores muito importantes relacionados com a eficiência da indexação de um documento e conseqüentemente, influenciam na revocação e na precisão. Esses fatores são a exaustividade e a especificidade. A Exaustividade define o número de diferentes conceitos (tópicos) que estão indexados e a especificidade define o grau de precisão da linguagem de indexação em descrever um dado documento (DANTAS, 2002).

Para o cálculo da precisão (precision) e revocação (recall), são utilizados as seguintes fórmulas matemáticas:

$$recall = \frac{Numero\,de\ documentos\ relevantes\ recuperados}{Número\,total\ \ de\ documentos\ relevantes}$$

$$precision = \frac{Número\,de\ documentos\ relevantes\ recuperados}{Número\,total\ \ de\ documentos\ recuperados}$$

Para avaliar o que afirmam os autores Santos e Silva (s.d.), realizamos um experimento para comprovar ou não esta afirmativa. Este experimento, foi conduzido conforme a metodologia adotada pelos autores SHAFI e RATHER (2005) presente no artigo “Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology”. Os termos utilizados na busca foram syndrome of down. Veja os resultados na tabela 16.

TABELA 16 - Precisão X Revocação

Sistema	Total de resultados	Total de resultados utilizados	Resultados iguais	Resultados relevantes	Resultados irrelevantes	Precisão	Recall
Altavista	4.920.000	15	9=yahoo 8=mamma 8=google 6=dmoz	13	1 site pessoal e 1 site repetido	13/15= 0,86	13 / 15+9+8+8+6 = 0,32
Google	3.370.000	15	6= metacrawler 8=altavista 10=yahoo 7=dmoz 3=mamma	14 (o único que recuperou a base Medline Plus)	1 site pessoal	14/15= 0,93	14 / 14+6+8+10+7+3 = 0,29
DMOZ	158	15	7=mamma 7=google 6=yahoo 6=altavista	14	1 site de outra doença, apenas cita a sindrome	14/15= 0,93	14/ 14+7+7+6+6= 0,35
Yahoo!	87	15	6= metacrawler 6=dmoz 9=altavista 10=google	14	1 (site não encontrado)	14/15= 0,93	14 / 14+6+6+9+10 = 0,31
Metacrawler	72	15	6=yahoo 6=google 4=altavista	12	3 (um site sobre jogos, um sobre venda de produtos de diversos tipos e outro sobre a venda de um software educativo	12/15= 0,8	12 / 12+6+6+4 = 0,42
Mamma	67	15	7=dmoz 9=yahoo 8=altavista 3=google	15	0	15/15= 1	15/ 15+7=9+8+3 = 0,35

Através dos resultados podemos verificar que o metabuscador “mamma” obteve o maior índice de precisão que todos os outros sistemas, inclusive diretórios. Isso significa que na busca realizada ele foi o sistema que obteve as respostas mais relevantes a nossa questão. Em relação a revocação ele obteve, aproximadamente, a média entre os outros sistemas.

Já o metacrawler, outro metabuscador, obteve o maior índice de revocação entre todos os sistemas e o menor índice de precisão. Ou seja, ele obteve mais

resultados que os outros, e mais resultados irrelevantes que os outros.

O google obteve um baixo índice de revocação (0,29) e um alto índice de precisão (0,93). Foi o único que recuperou documentos pertencentes à base Medline, uma importante fonte de informação em ciências da saúde. O que demonstra sua eficácia como mecanismo de busca é a comparação com o Altavista, mecanismo muito popular na web. O altavista recuperou maior quantidade e menor qualidade de resultados. O curioso nos resultados apresentados pelo Google, é o fato dele ter tido o mesmo índice de precisão que os 2 diretórios investigados. É curioso, pois a indexação dos diretórios é feita pelos humanos e no google é feita por máquinas. Talvez esse seja realmente o grande segredo do sucesso do Google, seu algoritmo diferenciado.

Nos diretórios, DMOZ e Yahoo, obteve-se o mesmo índice de precisão. A diferença esta na revocação. O DMOZ teve uma maior revocação. Talvez isso se justifique pelo fato dele possuir uma base de dados superior, conseqüência do maior número de editores em todo o mundo.

TABELA 17 - Resultados iguais

Altavista	Google	DMOZ	Yahoo	Metacrawler	Mamma
	8= altavista	6= altavista	9= altavista	4= altavista	8= altavista
9= yahoo	10= yahoo	6= yahoo		6= yahoo	9= yahoo
8= Mamma	3= mamma	7= mamma			
8= google		7= google	10= google	6= google	3= google
6= dmoz	7= dmoz		6= dmoz		7= dmoz
	6= metacrawler		6= Metacrawler		

Através da tabela 17 podemos perceber que existe uma grande sobreposição de resultados recuperados. Podemos notar também que o Google é o único que possui resultados iguais a todos os outros cinco sistemas. Os dois metabuscadores foram os sistemas que obtiveram as menores taxas de resultados iguais. Podemos justificar isso pelo fato de que os dois metabuscadores não fazem buscas em todos

os sistemas analisados neste trabalho. Dos analisados o Metacrawler faz buscas nas bases de dados dos Google e do yahoo, exatamente onde os resultados foram iguais. Já o Mamma, que mostrou um melhor desempenho na maioria dos experimentos feitos neste trabalho, faz buscas, entre outros, no DMOZ, Google e Yahoo. No entanto ele retornou resultados iguais ao Altavista, Google e DMOZ. Com estes resultados podemos perceber que não apenas os metabuscadores recuperam resultados iguais a outros sistemas, os mecanismos e os diretórios também fazem isso.

5 CONCLUSÃO

O visível crescimento no volume de informações disponibilizadas na web, proporcionadas pelos grandes avanços tecnológicos dos últimos tempos vem aumentando o desenvolvimento de projetos e pesquisas ligadas ao aperfeiçoamento de sistemas de busca na world wide web. Neste trabalho foram estudadas as três categorias de sistemas de busca; os mecanismos, diretórios e metabuscadores. Foram expostos seus conceitos, características, diferenças e semelhanças. Além de identificados e descritos seus principais representantes, foram realizados alguns experimentos acerca das ferramentas que cada sistema possui.

A maior dificuldade encontrada na organização da web, é o fato dela não ser organizada. As informações relevantes estão espalhadas e escondidas entre milhares de informações irrelevantes, o lixo eletrônico. Os sistemas de busca, que têm o objetivo de disponibilizar estas informações relevantes em fração de segundos, muitas vezes se tornam parte desta bagunça. Isso ocorre, pois nem sempre e não pouco freqüente, os sistemas não conseguem atender nem às questões que eles mesmos se propõem.

Podemos observar isto no caso dos metabuscadores que, como mostrado na tabela 9, obteve os mais baixos níveis na quantidade de resultados, o que contradiz seu maior objetivo, sua razão de existir, que é: uma maior cobertura de páginas na web. Talvez isso seja proveniente de problemas na tradução das buscas. Por isso, devemos ter cuidado na utilização dos metabuscadores quando a busca é feita com o uso de operadores booleanos, termos complexos ou que necessitem de qualquer ferramenta disponível nos mecanismos de busca.

Devido a esses fatores, é aconselhável o uso dos metabuscadores somente quando a busca é de caráter simples, ou seja, não precise de operadores e filtros.

No caso de buscas com termos simples, o desempenho é satisfatório. Comprovamos isso na tabela 16 (Precisão e Revocação) onde um dos metabuscadores, o “Mamma”, obteve o melhor índice de precisão comparado a todos os outros sistemas. Incluindo diretórios e mecanismos de busca.

Os mecanismos de busca são extremamente rápidos na recuperação de informações e, ao contrário dos metabuscadores, alcançaram em todos os experimentos um maior número (quantidade) de resultados. Por isso é interessante que sejam utilizados para recuperar informações provenientes de buscas complexas, pois as buscas simples tendem a recuperar uma enorme quantidade de respostas irrelevantes. A maior limitação dos mecanismos pode estar no fato de que, apesar de toda a sua tecnologia, são incapazes de perceber o significado das palavras.

Nesse sentido, é aconselhável que, sempre que possível, sejam utilizados mais de um termo de consulta com o intuito de esclarecer aos mecanismos o sentido/significado do termo pesquisado. Verificamos este fato, no experimento realizado através de palavras homógrafas. Além disso, a eficácia dos motores de busca está diretamente relacionada com a representação da busca, ou seja, com a precisão dos termos adotados na delimitação da pesquisa. Cabe ressaltar porem que a precisão dos termos que fazem parte das estratégias são escolhas do usuário e por isso demandam trabalho intelectual e conhecimento acerca das ferramentas de cada mecanismo. Como suas bases de dados são criadas através de uma indexação automática, indexam muito mais rápido do que os diretórios que possuem indexação manual. Por isso apresentam, freqüentemente, resultados compostos de dados mais atuais que os diretórios. Fator que faz diferença na busca de documentos científicos. O Mecanismo de busca que obteve os melhores resultados nas experimentações apresentadas neste trabalho foi o Google.

Já os diretórios possuem uma excelente proposta, baseada na organização hierárquica de assuntos. No entanto, essa vantagem não supera a desvantagem que vem junto com a demora na atualização da informação. A maioria das pessoas que utiliza os sistemas de busca para a recuperação da informação científica deseja a informação mais atual possível e os diretórios não conseguem suprir essa necessidade.

Alguns diretórios levam mais de dois anos para avaliar um site e inclui-lo em sua base de dados (neste tempo o site pode até não existir mais). Para muitas áreas do conhecimento, dois anos são necessários para que o conteúdo da informação não tenha mais valor prático. Existem diretórios como o Yahoo, por exemplo, que quando consultados sobre um tema que não possui, transfere a busca para outros sistemas, o Google no caso do yahoo, com a finalidade de que o pesquisador não fique sem resposta. Este é um cuidado que deve ser reconhecido.

É o mesmo cuidado que o bibliotecário deve possuir quando indica ao seu usuário onde ele pode encontrar certo documento, vendo que sua unidade de informação não o possui. Podemos pensar também, que a ordenação dos resultados apresentados em cada categoria deveria estar organizada, além da ordem hierárquica de assunto, por ordem decrescente de qualidade ou relevância. Isso evitaria que o pesquisador tivesse de percorrer todo o conteúdo da categoria para ter a certeza de que não deixou para trás nenhum site interessante. O Yahoo possui esta ferramenta. Acima dos resultados podemos optar por colocar os resultados em ordem de relevância ou não. No entanto, são raros os diretórios que adotam este critério de ordenação, podemos considerar o Yahoo uma exceção.

O Yahoo também foi o diretório que obteve os melhores resultados nos experimentos feitos neste trabalho, apesar do DMOZ possuir uma base de dados superior, em relação ao tamanho. Na hora do usuário optar por fazer sua busca em

um diretório, ele deve estar atento as suas limitações, tanto de atualidade como de quantidade. Apesar disso os diretórios são indicados para usuários iniciantes, o que pode proporcionar um conhecimento acerca do funcionamento da busca na web, trazendo conhecimentos para que ele arrisque suas buscas em sistemas mais complexos, como os mecanismos.

Qualquer que seja o sistema escolhido para uma determinada busca, o que importa, é que se conheça seu funcionamento e suas características, para que os resultados sejam satisfatórios e para que o usuário não desperdice seu tempo analisando resultados inúteis.

REFERÊNCIAS

ABREU, Marcelo Faoro de. **ALOI:** um agente para a localização e organização de informações**.** Porto Alegre, 2003. 82 p. Dissertação (mestrado em computação). Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, PPGC da UFRGS.

AGUILLO, I. F. **Documentación del curso Tratamiento documental de la World Wide Web**: técnicas de indización y clasificación de recursos en Internet. Disponível em: <www.cederul.unizar.es/noticias/sicoderxiii/po06.htm>. Acesso em: 14 mar. 2005.

AIRES, Rachel Virgínea Xavier; ALUÍSIO, Sandra Maria. Como Incrementar a Qualidade dos Resultados das Maquinas de Busca: da análise de logs à interação em português. **Ciência da Informação**, Brasília, v.32, n.1, p.5-16, jan./abr.2003. Disponível em: <http://www.ibict.br/cienciadainformacao/viewarticle.php?id=149&layout=abstract>. Acesso em: 26 ago. 2005.

ALENCAR, Maria Simone Menezes de. **Mecanismos de busca na Web:** uma análise da metodologia de estudos comparados. 2001. Dissertação (Mestrado em Ciência da Informação) – UFRJ/ECO - MCT/IBICT, Rio de Janeiro. Orientador: Maria de Nazaré Freitas Pereira.

ANDRADE, Carlos Alberto Valente. **Arcabouço para o desenvolvimento de portais colaborativos**. 2004. Instituto de Pesquisas Tecnológicas do Estado de São Paulo, São Paulo,. (Mestrado Profissional em Engenharia de Computação). Orientador Prof. Dr. Mário Yoshikazu Miyake, 96p.

BLATTMANN, Ursula; FACHIN Gleisy R. B.; RADOS, e Gregório J. Varvakis. Recuperar a informação eletrônica pela internet. Revista da ACB: Biblioteconomia em Santa Catarina, v. 4, n. 4, p. 9-27, 1999.

BRANSK, Regina Meyer. **Localização de informações na Internet**: características e formas de funcionamento dos mecanismos de busca. Texto preparado para o Curso de Extensão “Desenvolvimento de Negócios com o auxilio da Internet” do Instituto de Economia da Universidade Estadual de Campinas. Disponível em: <http://www.eco.unicamp.br/cefi/localizacao.doc>. Acesso em: 20 dez. 2004.

CARDOSO, Jiani Cordeiro; OLIVEIRA, João Batista. Problemáticas em interfaces de busca de bibliotecas digitais. s.d. Disponível em: <http://www.c5.cl/ieinvestiga/actas/ribie2000/papers/263/>. Acesso em: 10 out. 2005.

CASTRO, Maria Alice Soares de. **O Que é a Internet.** São Paulo, 2003. Disponível em: <http://www.icmc.usp.br/ensino/material/html/internet.html>. Acesso em: 13 jul. 2005.

_____. **O Que é World Wide Web.** São Paulo, 2003. Disponível em: <http://www.icmc.usp.br/ensino/material/html/www.html>. Acesso em: 14 jul. 2005.

CENDÓN, Beatriz Valadares. Ferramentas de busca na Web. **Ciência da Informação**, Brasília, v.30, n.1, p.39-49, jan./abr.2001. Disponível em: <http://www.ibict.br/cienciadainformacao/viewissue.php?id=17>. Acesso em: 13 jul. 2005.

DANTAS, Suzana. **Introdução a recuperação da informação**. Recife: Universidade Salgado de Oliveira, 2002. Disponível em: <www.di.ufpe.br/~sfd/universo/internet/resumo_ri.doc>. Acesso em: 17 out. 2005.

DETERS, Janice Inês; ADAIME, Silsomar Flôres. **Um estudo comparativo dos sistemas de busca na web.** Anais do V Encontro de Estudantes de Informática do Tocantins. Palmas, TO. outubro, 2003. p. 189-200. Disponível em: <http://www.ulbra-to.br/ensino/43020/artigos/anais2003/>. Acesso em: 02 maio 2005.

DOMINGUEZ, Adelaida Delgado. **Herramientas de búsqueda para la WWW**. CIVE2001 Congresso Internacional Virtual de Educação. Abril, 2001. Disponível em: <http://www.cibereduca.com/temames/ponencias/sept/p127/p127.htm#booleano>. Acesso em: 18 jul. 2005.

GONZALEZ, Marco Antonio Insaurriaga. **Recuperação da Informação e Processamento da Linguagem Natural**. Porto Alegre, 2002. 83 p. Exame de Qualificação (Doutorado em Ciências da Computação). Pontifícia Universidade Católica do RS. Programa de Pós-Graduação em Ciências da Computação, Faculdade de Informática da PUCRS. Disponível em: <http://www.inf.pucrs.br/~gonzalez/docs/qualificacao.pdf>. Acesso em: 03 ago. 2005.

LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência da Informação**, Brasília, v.31, n.2, p.60-71, maio/ago. 2002. Disponível em: <http://www.ibict.br/cienciadainformacao/viewissue.php?id=21> Acesso em: 06 ago. 2005.

MARCONDES, Carlos Henrique; GOMES, Sandra Lúcia Rebel. **O impacto da Internet nas bibliotecas brasileiras.** Disponível em: <http://www.rits.org.br/rets/re_editorial.cfm>. Acesso em: 20 jun. 2005.

MOTA, Davide. **Pesquisa na Internet**. Rio de janeiro: Senac Nacional, 1998. 128 p.

MOURA, Gevilacio Aguiar Coêlho de. **Sistemas de busca da web: diretórios e mecanismos de busca**. 2001. Disponível em: <http://www.quatrocantos.com/tec_web/sist_busca/index.htm>. Acesso em: 15 out. 2004.

NAHUZ, Fernanda. World Wide Web: aspectos teóricos dos mecanismos de busca. **Informação e Sociedade: estudos**, João Pessoa, v.9, n.2, p.1-7. Disponível em: <http://www.informacaoesociedade.ufpb.br/issuev9n299.htm> Acesso em: 15 jun. 2005.

NOVO, Maria do Carmo de Salvo Soares; NOVO, José Polese Soares. Internet Procurando uma Agulha no Palheiro Digital: o uso de mecanismos de busca. **Soc.**

Bras. da Ciência das Plantas Daninhas. Campinas, v. 5, n. 1, p. 20-27, 1999. Disponível em: <http://sites.mpc.com.br/jpsnovo/artigos/busca/>. Acesso em: 12 maio 2005.

ROSAS, João Luís; HOURMANT, Roger. **Dos motores de pesquisa...**. 2001. Disponível em: <http://users.skynet.be/penso.logo.encontro/curso/052_programas_de_metapesquisa.htm>. Acesso em: 13 set. 2005.

SANTOS, Ana Rosa dos; SILVA, Maria Conceição da. Caminhos para pesquisa na internet, nas áreas de nutrição e odontologia. Niterói: Universidade Federal Fluminense, s.d.. Disponível em: <http://www.ndc.uff.br/textos/14.a.pdf>. Acesso em: 17 out. 2005.

SANTOS, Eduardo Toledo; NASCIMENTO, Luiz Antonio do. Recuperação de informação em sistemas de informações na construção civil: o caso das extranets de projeto. In. _____. **Seminário de Tecnologia da Informação e Comunicação.** São Paulo: USP, s.d..

SANTOS, Gildenir Carolino; PASSOS, Rosemary. **Oficina II:** ferramentas e formas de pesquisa bibliográfica na internet. campinas, 2001. Disponível em: <http://www.bibli.fae.unicamp.br/edunet.ppt>. Acesso em: 15 out. 2004.

SEGREDOS do Google: desvende os recursos não revelados do poderoso sistema de busca!. São Paulo: Digerati Books, 2004. 94p.

SHAFI, S. M.; RATHER, R. A. Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. **Webology**, v. 2, n. 2, 2005. [Article 12]. Disponível em: <http://www.webology.ir/2005/v2n2/a12.html.> Acesso em: 11 out. 2005.

SILVEIRA, M. **Web Marketing: usando ferramentas de busca.** São Paulo: Novatec, 2002.

SOUZA, Renato Rocha; ALVARENGA, Lídia. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v.33, n.1, p. 132-141, jan./abr. 2004. Disponível em: <http://www.ibict.br/cienciadainformacao/viewissue.php?id=5>. Acesso em: 25 jul. 2005.

TEIXEIRA, Cenidalva Miranda de Sousa; SCHIEL, Ulrich. A Internet e seu Impacto nos Processos de Recuperação da Informação. **Ciência da Informação**, Brasília, v.26, n.1, p.1-14, jan./abr.1997. Disponível em: <http://www.ibict.br/cienciadainformacao/viewissue.php?id=29>. Acesso em: 03 ago. 2005.

URBANO, Magno. Guia de Referência. s. l., s. d. Disponível em: <http://www.fca.pt/livros-html/477_6.html#topicos>. Acesso em: 10 ago. 2005.

VIDOTTI, S. A. B. B.; SANCHES, S. A. S. **Arquitetura da Informação em Web Sites.** Campinas, 2004. Trabalho apresentado no Simpósio Internacional de Bibliotecas Digitais. Disponível em:

<http://libdigi.unicamp.br/document/?code=8302>. Acesso em: 06 jun. 2005.

WIKIPÉDIA: a enciclopédia livre. Open Directory Project. 2005. Disponível em: <http://pt.wikipedia.org/wiki/Open_Directory_Project>. Acesso em: 12 set. 2005.