WORLD WIDE WEB: ASPECTOS TEÓRICOS DOS MECANISMOS DE BUSCA

WORLD WIDE WEB: THEORETICAL ASPECTS OF SEARCH MECHANISMS

Fernanda **Nahuz**[1]

Resumo

Análise do uso de mecanismos de busca na WEB. São destacados alguns aspectos destes mecanismos, tais como: classificação, estrutura, recuperação, características e limitações. Com o advento da Internet surgiu a *World Wide Web* (Web ou WWW), que pode ser conceituada como um ambiente virtual de busca de informações do mundo da mídia eletrônica. A navegação por páginas iniciais *(homepages)* de sítios *(sites)* na WEB, tem colaborado no crescimento rápido da Internet. A disponibilização de documentos, informações, transações, comunicações em multimídia são operações conduzidas por navegadores *(browsers)*, tais como *Explore e Netscape*, que por sua vez, funcionam através de Mecanismos de Busca WWW *(Search Engines)* e instrumentam o cenário digital que emerge dos meios eletrônicos para as telas, monitores dos computadores, interligados pelo *Transmission Control Protocol / Internet Protocol (TCP/IP).*

Palavras Chave

INTERNET
MECANISMOS DE BUSCA
WORLD WIDE WEB

1 INTRODUÇÃO

A Internet, a mais vasta rede que conecta muitas outras espalhadas por mais de 170 países no mundo, tem em sua estrutura básica computadores interligados, que podem ser de diferentes tipos, portes, e sistemas, devendo apenas possuir o mesmo programa *Transmission Control Protocol / Internet Protocol* (TCP/IP). Para a navegação dos computadores interligados por TCP/IP moverem-se de um ao outro remotamente, entre os vários vínculos (*links*) da *World Wide Web*, os programas disponíveis são: *Mosaic, Netscape e Internet Explorer.*

A navegação por páginas iniciais *(homepages)* de sítios *(sites)*, a *Web* ou WWW, tem colaborado efetivamente na dimensão do crescimento da Internet. A disponibiblização de documentos, informações, transações, comunicações e multimídia na WWW conduzidas por navegadores (*browsers*), *Internet Explore* e *Netscape*, são localizadas por mecanismos de busca que instrumentam o cenário digital que emerge dos meios eletrônicos para as telas, monitores dos computadores interligados pelo protocolo TCP/IP. Por sua importância e características próprias, os mecanismo de busca WWW, são notadamente objetos de estudo nas áreas de conhecimento da informática, da comunicação, da lingüística, da psicologia, da ergonomia, da engenharia de redes, da ciência da informação.

[1] Aluna do Curso de Mestrado do *Departamento de Ciência da Informação e Documentação da Universidade de Brasília* (CID/ UnB). *E-mail:* fnahuz@unb.br

Os usuários da Internet têm o acesso à informação sem fronteiras, bem como a falta de normas e regulamentos na disponibilização de arquivos dos sítios W*eb* (*web sites*). O auxílio dos mecanismos da *World Wide Web* direcionam usuários na navegação pela grande teia (*Web*).

2 A INTERNET E A WORLD WIDE WEB

A *World Wide Web* trouxe a multimídia para computadores conectados a servidores *(hosts),* navegarem por sítios informacionais, interligados por vínculos *(links)* das páginas iniciais *(homepages)* de outros sítios, realizando a parte cliente, do sistema cliente-servidor que compõe a Internet. Dessa forma o *Web "permite o deslocamento de um documento, ou de um computador ao outro, por programas (softwares) denominados navegadores, capazes de possuírem texto, imagem, vídeo, e gráficos em multimídia*" (Notess, 1996).

As características marcantes da Web são a utilização: da hipermídia (som, imagem, hipertexto); do *Hypertext Markup Language* (html); do *hypertext transfer protocol* (http); dos programas gráficos para a navegação das páginas *Web* (*Mosaic, Netscape, Internet Explorer*).

3 MECANISMOS DE BUSCA DA WORLD WIDE WEB

São mecanismos de busca utilizados na recuperação de arquivos representados na Internet e que têm sua **estrutura** composta por três elementos básicos:

a) Dispositivo de Colheita: robôs e/ou rastrejadores, que obtêm parte ou toda informação do servidor de uma rede de aplicações como o *File Transfer Protocol* (FTP) e a *Web.*

b) Mecanismo de Indexação: organizam e atualizam dados amontoados de um processo de colheita. O produto final será uma base de dados, um catálogo ou um diretório de assuntos.

c) Sistema de Busca: permite vários usuários realizarem buscas com facilidade em uma determinada Base de Dados de várias formas, dentre elas a busca booleana, truncagem, ou ainda a busca pelo próprio http do sítio.

Tais mecanismos estão **classificados** como:

a) Mecanismo na busca: busca direta (*search engines*); busca indireta (*catalogs ou subject directory*).

b) Aplicação: TELNET, FTP, *World Wide Web, Gopher, WAIS*. Esta classificação por aplicação está representada na tabela a seguir (Chu, 1997, p.24).

APLICAÇÃO	**BUSCA DIRETA** (search engine)	**BUSCA INDIRETA** (catalog/subject directory)
TELNET	(diversas)	Hytelnet
FTP	Archie	-
WAIS	(embutida)	(via gopher)
GOPHER	Veronica, Jughead	Gopher Jewls
WWW	Alta Vista, etc	Yahoo, etc

c) Conteúdo: cobertura exaustiva (*excite***,** *lycos*); assunto específicos; recuperação de tipos específicos de informação; *Switchboard* - endereço de números de telefones; *Four 11, Whowhere* - endereços eletrônicos (e-mail); *Mapblast* - mapa detalhado para endereços.

As **técnicas de recuperação**, que são utilizadas pelos mecanismos de busca, segundo Chu (1997), Notess (1996) e Westera (1996) estão relacionadas abaixo:

a) Busca Booleana *(Boolean Search)*

Aplicada a todos os Mecanismos de Busca com exceção do *Archie*. Os símbolos booleanos (+, -), ao contrário dos operadores (*and, and not, or*), são usados para realizar facilmente a busca booleana, não havendo problemas por parte dos usuários. Por exemplo, no AltaVista, o símbolo + significa *and*, e o símbolo - significa *not*, enquanto a ausência do uso de símbolos representará a união através do operador *or*. Em outros casos os operadores booleanos estarão apresentados como ítem nos menus. Entretanto, *Lycos* utiliza *match all terms* para indicar *and*, e *match any terms* para *or*.

b) Truncagem ou Máscara (*Truncation*)

É um meio de re-chamada improvisória, muitas vezes utilizada pelos mecanismos na Internet. Alguns mecanismos truncam automaticamente, como o *Lycos*; enquanto que outros a executam apenas em certas circunstâncias. A maioria dos mecanismos de busca, *truncam* a direita dos termos; como se fossem sufixos dos mesmos; poucos permitem truncagem no meio. O *Archie***,** realiza sua truncagem a esquerda do termo, raramente vista na recuperação de informação na Internet, segundo Felt (1996) e Chu (1997).

c) Busca Por Proximidade (*Proximity Search*)

Na Internet**,** a busca por proximidade é empregada em escala limitada em apenas alguns mecanismos (ex: *Infoseek* e Alta Vista) utilizam o operador *Near*. Outros operadores de proximidade (ex: *Near*) não podem ser utilizados em buscas de rede. A busca por proximidade é estritamente falada como uma expressão de busca no sistema de informação de uma rede. Todavia, um grande número de mecanismos de busca da Internet não possuem esta capacidade, que é essencial na precisão para a recuperação da informação em rede.

d) Busca Por Frase (*Concept Search*)

A recuperação busca por frase, destina-se a informação relacionada a um conceito dado. Por exemplo; em uma busca onde a expressão é transporte público, informações que tratem de ônibus e metrô deverão ser recuperadas por mecanismos, como o *Excite* e o *Infoseek*, caracterizadas principalmente por esta capacidade.

Limitações dos mecanismos

a) Manipulação de Cenários

Aplicada aos mecanismos que possuem em seu menu de busca a opção *Revise Search*, que permite aos seus usuários adicionar ou eliminar termos para uma busca, a exemplo o *HotBot*.

b) Busca Limitada

A busca limitada apresenta-se de diversas formas na Internet. Alguns mecanismos limitam por data (Alta Vista) e por lugares (*HotBot*).

c) Inexistência da Busca por Vocabulário Controlado

A busca por vocabulário controlado não é realizada na Internet, devido ao problema da alta revocação, e da baixa precisão nas realizações de busca. Controlar as linguagens será possível quando os dados na rede forem buscados, recuperados e indexados por vocabulários normalizados, como é feito tradicionalmente nas bases de dados bibliográficas, ou quando, a busca feita por intermédio da inteligência artificial, elaborar a recuperação e o estoque da informação.

d) Outras Limitações

São impotentes na remoção de duplicações e/ou de vínculos mortos nos resultados de suas buscas, ainda que possuam o campo *filtering search results* no menu, excetuando os mecanismos de busca múltiplos, como o *Metacrawler*.

Tipologia Dos Mecanismos De Busca WWW

Os mecanismos de busca da Internet *(Internet Search Tools)* são como grupos prontos, contidos em um amplo espaço de buscas facilitadas, apesar de serem recentes. Chu (1997):

a) *Subject Directory* (Busca Indireta)

Possuem índices organizados hierarquicamente; listas seletivas, com margem de erro humano; economizam tempo em buscas de informações irrelevantes; são conhecidos como *Catalogs* ou *Subject Directory*. Ex: *Yahoo*!

b) *Search Engine* (Busca Direta)

Realiza busca por termos ou expressões; possui base de dados criada automáticamente por programas; a solicitação e a colheita de dados são realizadas pelas visitas, rastrejamentos, dos robôs (programas); são conhecidos por *Search Engine.* Ex: Alta Vista .

c) Mecanismo de busca múltiplo

Realizam buscas em diferentes mecanismos, em uma única página; nem sempre são atualizados em suas buscas; realizam a colheita pelos últimos resultados das bases de dados, nos diferentes mecanismos recuperados; alguns destes mecanismos navegam apenas em computadores de mesa (*desktops*), por terem programas comerciais; são conhecidos por *crawlers, worms, spiders*. Ex: *Metacrawler, Miner*.

Segundo Slot (1996) os mecanismos de busca têm as seguintes **características**;

a) Atributo

Os mecanismos de busca WWW em informações gerais, podem ser: revisões de fontes *Web*; catálogo de busca; base de dados de busca; índice de busca; catálogo de assuntos.

b) Atualização

É muito importante estar atento ao período em que o mecanismo realiza a sua colheita e/ou rastrejamento, para evidenciar a atualização de sua base de dados. Ex: *Excite!NetReviews* - atualização diária

c) Acervo

Quantos documentos o mecanismo adiciona diariamente; Ex: *Lycos* - mais de 1.000 documentos diários. Possuia 20.000 vínculos *Web*, quando Slot (1996), realizou seu estudo.

d) Acesso

O acesso geralmente é gratuito. Ex: *Lycos*, AltaVista, *Yahoo* (a maioria). Registram-se poucos mecanismos que cobram por seus serviços.

Quadro demonstrativo dos critérios apresentados por Slot (1996)

MECANISMOS DE BUSCA WEB	DESCRIÇÃO	ATUALIZAÇÃO	ACERVO	ACESSO
BUSCA DIRETA	Ver tópico	1.3 dos Aspectos	Teóricos dos	Mecanismos de Busca na WWW
Alta Vista	Base de Dados de Informações Gerais	Corrente		

Infoseek Ultra	Catálogo de Informações Gerais	Corrente	Mais de 1 milhão de páginas *Web* (*Webpages*)	Busca livre através de bases de dados. Usuários cadastrados possuem opções adicionais mediante pagamento.
Infoseek Guide			10 mil *Usenet Groups* e outras numerosas fontes *online*	
Excite! *NetSearch*	Base de Dados de Informações Gerais	Corrente	Mais 50 mil *abstracts* de documentos *Web*	Acesso livre e ilimitado
Webcrawler	Índice de Informações Gerais	Corrente	Mais de 420 mil documentos *Web*	Livre e ilimitado
BUSCA INDIRETA	Ver tópico	1.3 dos Aspectos	Teóricos dos	Mecanismos de Busca na WWW
SUBJECT DIRECTORIES				
Yahoo!	Catálogo de Assuntos de Informações Gerais	Corrente	Mais de 370 mil documentos *Web* e *Gopher*	Livre e ilimitado
Excite!*NetReviews*	Revisões de Fontes Gerais da *Web*	Corrente	Mais de 50 mil documentos *Web*	Livre e ilimitado

Para Zif-Davis (1997) habilidade de navegação (*browseability*) refere-se a legibilidade e velocidade do navegador nos elementos abaixo:

- A relevância das buscas realizadas;
Ex: *Lycos* / *Infoseek* - são os favoritos.

- O número de vínculos recuperados;
Ex: *Webcrawler* / WWW *Worm* / *Internet Exploration Page*

- Os resultados das buscas com páginas iniciais ativas, sem mensagem de erro.
Ex: *Yahoo*! / *Infoseek*

Arents (1997), utilizando-se da classificação apresentada no quadro abaixo, adiciona mais dois critérios em relação aos mecanismos de busca WWW:

a) Usabilidade - Quão fácil é entender, formular e submeter questões *(queries)* aos mecanismos de busca *Web*.
b) Eficácia - Refere-se a quantidade, precisão e legibilidade dos resultados que retornam das buscas realizadas nos mecanismos WWW.

CLASSIFICAÇÃO	USABILIDADE / EFICÁCIA
THE BEST	(4) estrelas douradas
VERY GOOD	(3) estrelas douradas (1) prateada
GOOD	(2) estrelas douradas (2) pratedas
ÚTIL	(1) estrela dourada (3) prateadas
SEM CLASSIFICAÇÃO	(4) estrelas prateadas

4 ALGUNS MECANISMOS DE BUSCA WWW

BUSCA DIRETA (*search engines*)	LOCALIZA UM DOCUMENTO BASEADO EM CONTEÚDOS.
Alta Vista	- Reivindica ser capaz de indexar e buscar páginas *Web* e *Usenet Newsgroups* dez vezes mais rápido que outras *search engines*. - *Search Simple* / *Expert Search* / *Background Information*.
Infoseek Ultra	- Única *search engine* que retira vínculos duplicados e mortos quando realiza a sua indexação. - *Search Simple* / *Expert Search* / *Background Information*
Infoseek Guide	- Único serviço que integra totalmente a navegação do *Internet Directory* com o *Internet Searching*.
Excite!*NetSearch*	- Base de dados de mais de 1 milhão de páginas *Web* e deixa por duas semanas as *Usenet News* e os *Classified Advertisements*.
Webcrawler	- Opera por atravessador na *Web* e também constrói um índice para uso posterior, ou por busca em tempo real.
BUSCA INDIRETA (*subject directories*)	LOCALIZA UM SÍTIO NA WEB BASEADO NO QUE SE TRATA.
Yahoo!	- O diretório da Internet. Ainda é o meio mais popular de localizar um sítio na *Web*. - *Search Simple* / *Expert Search* / *Background Information*.
Excite!*NetReviews*	- Uma coleção de sítios *Web* e *Usenet Groups* revistos organizada hierarquicamente. - *Search Simple* / *Expert Search* / Background *Information*.

Abstract

Analyses the use of search mechanisms in the WEB. Some aspects of these mechanisms are highlighted, such as: classification, recuperation, characteristics and limitations. World Wide Web (Web or WWW) arose with the advent of Internet which can be conceptualised as a virtual environment for information in the world of the electronic media. Navigation by homepages of the sites on the WEB has contributed to the rapid growth of Internet. The availability of documents, information, transactions, communication in multimedia are operations conducted by browsers such as Explore and Netscape which in turn function by means of Search Engines and instrumentalise the digital scenario which emerges from the electronic media to the screens, monitors of computers, linked together by the Transmission Control Protocol / Internet Protocol (TCP/IP).

Keywords

INTERNET
SEARCH MECHANISMS
WORLD WIDE WEB

REFERÊNCIAS BIBLIOGRÁFICAS

ARENTS, Hans C. ***A selection of Internet search tools***. [online]. Available from World Wide Web: <http://www.mtm.kuleuven.ac.be/Services/search.html>. [out. 1996].

BIRMINGHAM, Judy. ***Internet search engines***. [online]. Available from World Wide Web: <http://www.stark.k12.oh.us/Docs/search/header.html>.

CHU, Heting. Internet search tools: what can they offer to users? - database access. In: NATIONAL ONLINE MEETING, 18, 1997, New York. ***Anais...*** Medford: Information Today, 1997. p.73-80.

_____. ROSENTHAL, Marilyn . Search engines for the World Wide Web: a comparative study and evaluation methodology. [online]. In: ASIS 1996 ANNUAL CONFERENCE PROCEEDINGS. 1996, New York. ***Anais...*** New York: ASIS, 1996. Available from World Wide Web: <http://www.asis.org/annual-96/Eletronic Proceedings/chu.html>.

FELT, Elizabeth. ***Analysis of web robots***. [online]. Available from World Wide Web: <http://www.wsulibs.wsu.edu/general/robots.htm.>

KOCH, Traugott. ***Internet search services***. [online]. Available from World Wide Web: <http://www.ub2.lu.se/tk/demos/DO9603-meng.html>. [22 abr. 1996].

LIU, Jian. ***Understanding WWW search tools***. [online]. Available from World Wide Web: <http://www.indiana.edu/~librcsd/search/>.

NOTESS,Greg . ***Searching the Internet?*** learn the tricks of the trade. [online].. Available from World Wide Web: <http://www.imt.net/~notess/search/about.html>. [12 set. 1997]

SLOT, Matt. ***The matrix of Internet catalogs and search engines: a comparison of Internet indexing tools***. [online]. Available from World Wide Web: <http://www.ambrosiasw.com/~fprefect/matrix/matrix.html>. [29 set. 1997].

ZIFF-DAVIS. ***Hide and go seek***. [online]. Available from World Wide Web: <http://www.zdnet.com/pccomp/features/internet/search/sub1.html>.

WESTERA, Gillian . ***Robot-driven search evaluation overview***. [online]. Available from World Wide Web: <http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/>.

WHAT is the Internet? [online]. Available from World Wide Web: <http://www.delphi.com/navnet/faq/internet.html>.