Universidade de Brasília
Faculdade de Economia, Administração, Contabilidade e Ciência da Informação
Departamento de Ciência da Informação e Documentação

Dayana Ester Andrade Figuereido

Recuperação da informação: uma análise sobre os sistemas de busca da web

Brasília

2006

Dayana Ester Andrade Figuereido

Recuperação da informação: uma análise sobre os sistemas de busca da web

Orientadora:

Prof. Dr. Marisa Brascher Basílio Medeiros

Monografia apresentada ao Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito para obtenção do título de bacharel em Biblioteconomia.

Brasília

2006

F475r Figuereido, Dayana Ester Andrade
Recuperação da informação: uma análise sobre os sistemas de busca da web/ Dayana Ester Andrade Figuereido.-- Brasília: CID/UNB, 2006.
61 f. (Monografia de graduação)

1. Recuperação da Informação. 2. Sistemas de Busca. 3. Ciência da Informação 1. Título.

CDU 025.4.03

Dedico este trabalho a meus pais, meu marido Marcelo e a minha filha Luíza, pelo carinho, incentivo e paciência.

Agradecimentos

O agradecimento é uma forma de reconhecer que as dificuldades da vida não são superadas sozinhas. Cada fase desse caminho só foi possível de ser superada por que tive ao meu lado pessoas maravilhosas.

Assim, primeiramente agradeço a Deus por todas as dávidas e pela força que recebi.

Ao meu pai Severino Nascimento de Figuereido e à minha mãe Linastern B. Andrade Silva pelo exemplo de força e coragem.

Em especial a meu pai que, como sábio lutou com todas as forças e meios, mesmo passando por dificuldades financeiras, para que seus filhos pudessem sempre ter boa educação.

A meu marido Marcelo de Souza Veras e minha filha Luiza Andrade Veras pelo incentivo e principalmente por serem motivadores de muita alegria e satisfação.

Às professoras Simone Bastos que iniciou comigo essa jornada e Marisa Brascher que me ajudou a seguí-lo.

Aos amigos Daniela Galvão e Felipe Kenzo pelo incentivo e carinho.

À D. Antônia que com muita paciência cuidou de minha filha durante minhas ausências para que eu pudesse atingir meu objetivo.

E por fim aos colegas de trabalho do Conselho Federal de Contabilidade Lúcia Helena e Marcelo Santana Costa.

Resumo

Nos últimos anos a *we*b tem crescido de forma rápida e exponencial, dessa forma, os sistemas de busca surgem como uma tentativa de facilitar o acesso a esse grande volume de informação disponibilizada por meio da *web*, oferecendo ao usuário a possibilidade de recuperar informações que satisfaçam sua necessidade**.** O presente trabalho visa apresentar aos usuários da *web*, o cenário atual da recuperação da informação na *web* no que diz respeito aos sistemas de busca, diretórios e mecanismos de busca, e quais as suas tendências para o futuro.

Palavras-Chave: Recuperação da Informação, Mecanismos de Busca, Web.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

A *web* nos últimos anos passou por um processo de crescimento e popularização muito grande. Porém este crescimento não ocorreu de forma ordenada e controlada. A linguagem HTML, de fácil manuseio e sem padronização contribui para que diversos documentos sejam disponibilizados na *web* sem, no entanto, haver controle de conteúdo. Da mesma forma, o senso de urgência trazido pelos novos paradigmas da chamada Sociedade da Informação, provocou o desenvolvimento de páginas sem contemplar aspectos de padronização exigidos pela comunidade científica.

Os sistemas de busca surgem como uma tentativa de facilitar o acesso a esse grande volume de informação disponibilizada por meio da *web*, oferecendo ao usuário a possibilidade de recuperar informações que satisfaçam sua necessidade. Estes sistemas, após a indexação das páginas da *web*, comparam a informação solicitada pelo usuário com a que está contida em seu banco de dados e retornam ao usuário uma lista de documentos com informações similares a essa informação solicitada.

A eficiência de um processo de recuperação de informação está diretamente ligada à estratégia de busca elaborada pelo usuário. Dessa forma, o conhecimento dos sistemas busca, diretórios e mecanismos de busca, torna-se essencial para a elaboração de uma estratégia de busca eficaz e para o resultado proveitoso de uma pesquisa na *web*.

Face à importância desse conhecimento, o presente trabalho visa apresentar aos usuários da *web*, o cenário atual da recuperação da informação na *web* no que diz respeito aos sistemas de busca, diretórios e mecanismos de busca, e suas perspectivas para o futuro. Para atingir tal objetivo foi feita uma pesquisa de literatura e exploratória sobre os principais sistemas de buscas da *web* na atualidade, suas características, limitações, problemas e por fim os projetos de aprimoramento da *web*: *Web Semântica* e *Web Intelligence.*

2 Problema

O acelerado crescimento tecnológico e o surgimento de novos sistemas de busca impõem cada vez mais que os usuários da *web* e profissionais da informação mantenham-se atualizados frente a esses recursos.

Apesar dos esforços dos desenvolvedores em oferecer sistemas de busca com interfaces amigáveis com orientações por meio de *menu*s ou oferecendo recursos especiais para usuários inexperientes, a maioria dos usuários dos bancos de dados na *web*, não tem conhecimento de controles mais avançados, não sabe elaborar uma estratégia adequada de busca e não explora adequadamente todo o potencial dos sistemas de busca.

A motivação deste trabalho veio da percepção da escassez de literatura atualizada em língua portuguesa sobre os sistemas de busca, em linguagem acessível para usuários leigos em tecnologia da informação.

Dentro desta perspectiva o trabalho proposto visa responder a seguinte indagação:

Qual o cenário atual que envolve os sistemas de busca da *web* e suas perspectivas para o futuro?

3 Objetivos

3.1 Objetivo Geral

Oferecer uma visão geral das principais categorias de sistema de busca que a *web* dispõe na atualidade para recuperar informação e suas perspectivas futuras.

3.2 Objetivos específicos

- Discorrer o sobre o processo de recuperação da informação;
- Discorrer sobre as características dos principais tipos de sistemas de busca da *web*, apresentando as limitações, dificuldades e problemas enfrentados por estes sistemas;
- Apresentar os principais projetos de aprimoramento da *web*: Web *Semântica* e *Web Inteligence.*

4 Justificativa

Os benefícios do uso da *web* como forma de recuperar a informação e para transmissão do conhecimento produzido são enormes, entre eles pode-se citar: a rapidez e facilidade no acesso á informação, visão de diversos pontos de vista sobre determinado assunto, aumento do intercâmbio informacional e comunicação interpessoal entre outros. Os sistemas de busca possuem um papel fundamental na recuperação a informação na *web*, pois é por meio dos mesmos que o usuário busca e acessa as informações dispersas na rede.

O conhecimento sobre os sistemas de busca da *web* também possibilita ao usuário a utilização mais eficiente dos recursos disponibilizados pelos mesmos, além de trazer respostas mais relevantes e precisas às suas pesquisas. Nota-se claramente que os usuários se limitam a usar somente os recursos de busca básica sem saber sequer da existência de recursos da buscas mais avançados que tornam a pesquisa mais eficiente, como por exemplos os conectores *booleanos*.

Da mesma forma o profissional da informação como mediador entre o usuário e o sistema de recuperação de informação, deve estar informado e freqüentemente atualizado sobre os sistemas de busca para poder satisfazer a necessidade informacional de seu usuário.

Esta pesquisa ao apresentar o cenário atual dos sistemas de busca da *web*, procura oferecer um instrumento de apoio ao profissional da informação e aos usuários da *web* em geral e assim proporcionar alternativas de uso mais eficaz de sistemas e redes de informação na recuperação da informação da web.

Alem disso, justifica-se a realização deste trabalho devido à ausência de literatura recente sobre o tema em língua portuguesa. Numa pesquisa feita recentemente na base de dados do LIS- *E-prints in Library and Information Science* o documente mais recente encontrado que tratava especificamente de sistemas de busca era de 2003 em língua espanhola.

5 Metodologia

Para atingir o objetivo proposto, foi elaborada uma pesquisa bibliográfica sobre o tema desta monografia em fontes impressas e na *web*. Sua finalidade é colocar o pesquisador em contato direto com tudo aquilo que foi escrito sobre determinado assunto. (LAKATOS, 1986).

Inicialmente foi feito um levantamento bibliográfico sobre a recuperação informação no contexto da Ciência da Informação e sobre o processo de recuperação da informação. Em seguida foi elaborado um estudo sobre a criação da *web* desde o surgimento da Internet até a criação do Consórcio W3C.

Na segunda etapa foi elaborado um levantamento na literatura sobre os sistemas de busca da *web*. Nesta fase foram estudados os dois principais tipos de sistemas: os diretórios e os mecanismos de busca e ainda os metabuscadores. Por meio deste levantamento procurou-se identificar características, limitações, dificuldades e problemas comuns à maioria dos sistemas de busca da *web*. Foram selecionados os sistemas de busca mais populares e citados na literatura mais atual consultada e foi apresentada uma visão geral de suas características e estratégias de buscas.

Na terceira e última etapa, foi feita uma pesquisa na literatura e na *web*, sobre os principais projetos de aprimoramento da recuperação da *web*.

Os principais *sites* e base de dados consultados foram:

www.searchenginewatch.com- *Search Engine Watch*
www.searchengineshowdown.com- *Search Engine Showdown*
www.google.com.br- *Google*
www.scirus.com.br- *Scirus for Scientific Information*
www.w3.org- W3C *World Wide Web Consortium*
www.usp.br- Universidade de São Paulo
E-*prints in Library and Information Science*- www. eprints.rclis.org
Biblioteca Digital de Teses e Dissertações da UFRGS- www.teses.usp.br
Banco de Teses de Dissertações da UnB- www.bce.unb.br
LISA- *Library and Information Science Abstracts* (IBICT)

6 Recuperação da Informação

Antes de falarmos sobre a recuperação da informação na *web*, e de qualquer tecnologia que a envolva, é imprescindível que se fale dos conceitos de recuperação da informação dentro do campo da Ciência da Informação, já que os dois campos estão inter-relacionados e possuem basicamente, como será descrito a seguir, fundamentos idênticos.

6.1-Conceituação

Ao longo do período da Segunda Guerra Mundial (1939-45), ocorreu um aumento considerável de literatura científica resultante da demonstração de como a Ciência poderia ser utilizada como prática e benefícios para a Guerra. A necessidade de se obter novas informações científicas e técnicas, num curto espaço de tempo fizeram com que pesquisadores de diversas áreas desprendessem esforços para a criação e organização de serviços especiais de informação. Assim, recuperação da informação surge como uma possível solução para o problema de explosão informacional identificada por Bush em 1945, como sendo o irreprimível crescimento exponencial da informação e de seus registros, particularmente em ciência e tecnologia. No contexto da Ciência da Informação, o termo "recuperação da informação" é bastante diversificado.

De acordo com Calvin Mooers (1951 apud SARACEVIC, 1996, p. 44), o termo recuperar informação "engloba os aspectos intelectuais de descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação".

Belkin e Croft (1987) definem o processo de recuperação de informação como um processo de localização de e itens de informação que tenham sido objetos de armazenamento, com a finalidade de permitir o acesso dos usuários aos itens de informação, objetos de uma solicitação. A recuperação da informação se dá pela comparação do que se solicitou com o que está armazenado e com o conjunto de procedimentos que esse processo envolve.

Lancaster (1978), afirma que recuperação da informação é um termo sinônimo de busca de literatura sendo, portanto um processo de para se buscar uma coleção de documentos.

Robredo (2005), define a recuperação da informação como a finalidade do trabalho documentário que envolve os processo de seleção, aquisição, descrição bibliográfica, análise e indexação. Como resultado das operações realizadas no processo de busca pode-se selecionar documentos (ou suas referências) de potencial interesse.

Numa concepção mais abrangente, Bastos (1994) define a recuperação da informação como um processo de comunicação onde se relacionam emissor e receptor com a finalidade de descobrir uma necessidade de informação. Ao fazer uma pergunta ao sistema o homem funciona como o emissor e o sistema como receptor. Em contrapartida o sistema ao apresentar sua resposta passa a ser o emissor e o homem o receptor. Essa interação se torna viável através do uso da linguagem. Dessa forma o estudo do processo de recuperação é multidisciplinar, pois envolve conhecimentos lógicos, tecnológicos e lingüísticos.

Este trabalho optou por discorrer sobre os processos e recursos voltados para a recuperação e disseminação de informações no que diz respeito aos sistemas de busca da web. Os sistemas voltados para o tratamento da informação (catalogação, indexação, classificação) não foram abordados, embora complementar aos processos de busca de informação.

6.2 Sistemas de Recuperação de Informação

Nesta seção apresentaremos os conceitos, características e funcionalidade dos Sistemas de Recuperação de Informação dentro da perspectiva da Ciência da Informação com o objetivo de demonstrar o quanto os mesmos se assemelham com o objeto desse estudo: os de Sistemas de Buscas da *Web*.

Grande parte da literatura define os Sistemas de Recuperação da Informação como qualquer sistema automatizado que visa à recuperação da informação sejam eles, catálogos de bibliotecas ou as bases de dados sendo, portanto, um subconjunto dos Sistemas de Informação. (MACEDO, 2005). Esses últimos são definidos na literatura como os próprios serviços de informação, tais como bibliotecas ou centros de informação.

Os primeiros sistemas automatizados de recuperação de informação se desenvolveram a partir do surgimento dos computadores e têm suas modificações atreladas ao

desenvolvimento da tecnologia da informação, mais especificamente a capacidade de armazenamento e processamento dos computadores.

Kent (1972), no início da revolução dos computadores, define a recuperação da informação como um ato de investigar ou explorar com o fim de encontrar algo perdido utilizando qualquer processo de mecânico de gravação do conhecimento. Entretanto, Kent explica que processo mecânico apenas facilita o acesso para os futuros usuários e que a recuperação é na verdade a pesquisa dos papéis escritos realizada pelas máquinas.

Para Rowley (2002) os sistemas de recuperação da informação e computadores quase foram usados como sinônimos, porém antes do surgimento de qualquer computador e da própria informática os sistemas de fichas e arquivos baseados em papel já existiam. Para a autora, os sistemas podem ser compreendidos como se fossem formados por três etapas:

- indexação-que definida por Robredo (2005), consiste em indicar o conteúdo temático de uma unidade de informação, mediante a atribuição de um ou mais termos (ou código) ao documento, de forma a caracterizá-lo de forma unívoca;
- armazenamento-processo geralmente feito por meio de computadores que guardam arquivos de documentos, índices e as base de dados que contém os registros dos documentos representados.
- recuperação da informação-consiste em identificar, no conjunto de documentos (*corpus*), quais informações atendem à necessidade de informação do usuário.

A indexação tem como propósito principal, como explica Lancaster (2004), representar documentos publicados para que possam ser incluídos numa base de dados. Essa base de dados de representações pode ser impressa, em formato eletrônico, ou em fichas.

A eficiência de um sistema de informação está diretamente ligada á estratégia de busca formulada pelo usuário, à qualidade com que a indexação foi realizada, à qualidade do vocabulário controlado entre outros fatores. As estruturas de estratégia de buscas serão melhor descritas na seção 13.

A última etapa está diretamente ligada às duas etapas anteriores, influencia diretamente no modo de operação do sistema e deve estar sincronizada com qualquer modelo de busca e recuperação de informação que possa ser proposta como solução para as necessidades de informação do usuário.

Os sistemas de recuperação da informação lidam com a representação, armazenamento e acesso aos documentos originais (documentos eletrônicos) ou a representações desses documentos, como dados bibliográficos, catalográficos ou referenciais e tem como principal propósito facilitar a recuperação da informação, desta forma devem prover os mecanismos que possibilitem a busca, a seleção, à localização e o acesso às informações relevantes aos seus usuários.

Numa visão mais abrangente Robertson (1981) afirma que os sistemas de recuperação da informação são um conjunto de regras e procedimentos que executados a partir da ação humana e/ ou máquinas que engloba atividades de indexação, formulação de busca, busca, retroalimentação e construção da linguagem de indexação. Esse processo visa fornecer resposta para uma determinada demanda que satisfaça a necessidade uma específica de informação do usuário.

Completando a visão de Robertson (1981), Belkin (1981) afirma que os sistemas recuperação de informação lidam com conceitos que são a base para o processo de recuperação da informação são eles: necessidade de informação, desejo, informação, significada ou a falta do mesmo, satisfação do usuário e efetividade da informação.

Paradoxalmente, de acordo com Braga (1995) os Sistemas de Recuperação de Informação, só conseguem recuperar uma informação em potencial, uma probabilidade de informação, que só vai se consubstanciar, se também houver uma identificação (em vários níveis) da linguagem do documento e uma alteração, uma reordenação mental do receptor usuário.

6.3 Processo de Recuperação da Informação

O processo de busca e recuperação de informação consiste em localizar documentos e itens de informação que tenham sido armazenados. Em geral as informações são recuperadas das bases de dados através de expressões de busca que utilizam termos e operadores. Essa equação culmina na expressão de sua necessidade informacional. Corresponde ao processo de extração e síntese dos conceitos da demanda do usuário e na tradução destes conceitos em termos utilizados pelas bases de dados. (BASTOS, 1994).

Quantos mais termos combinados por meio das expressões de busca maior a chance de se recuperar os documentos, isto é, maior revocação, porém corre-se o risco de recuperar documentos que não satisfaçam a necessidade do usuário. Por outro lado, a seleção de termos e a associação de operadores lógicos aumentam a precisão da informação recuperada.

Os termos a serem escolhidos durante a elaboração de uma estratégia de buscas podem ser da linguagem natural (do próprio usuário) ou da linguagem da base de dados (linguagem documental). O êxito da consulta vai depender se o termo utilizado pelo usuário for o mesmo usado na base de dados para a representação do documento.

As representações dos documentos, como citado na seção anterior, são feitas por meio da indexação dos documentos e dos resumos.

A realização de uma busca de base de dados consiste numa sucessão de etapas, que conduzem a execução da pergunta e são descritos segundo Amat (1989) da seguinte forma:

- definição de uma pergunta, delimitando a necessidade de informação do usuário (objeto de busca) através de uma entrevista ou por meio de uma solicitação de busca que estabelece um perfil (individual ou coletivo) que descreve os temas pertinentes;

- análise e preparação da busca: se determinam os conceitos mais adequados tendo em conta as formas de expressá-los.

- tradução da pergunta na linguagem do sistema: uma vez obtida a lista de conceitos, se deve representá-los através de termos da linguagem do sistema que se interroga e eleger os operadores que vão estabelecer as relações entre eles;

- resposta da consulta: consiste em oferecer os documentos que vão responder a busca. Neste ponto o usuário fará sua avaliação, permitindo a reformulação, em caso de algum problema.

7 Recuperação da Informação na Web

7.1 Histórico do surgimento da Web

A história da *Web* está ligada à evolução da *internet* e por isso descreveremos primeiro a evolução da mesma. O surgimento da *internet* começa no período pós-guerra, no ano de 1957, quando os Estados Unidos criaram o Departamento de Defesa (DoD) e a ARPA (*Advanced Research Projects Agency*) em resposta ao sucesso do programa espacial soviético representado pelo lançamento do *Sputnik*, uma pequena esfera de alumínio de 84 quilos equipada com um transmissor contendo em seu interior a célebre cachorrinha chamada Laika.

A *internet* foi criada com objetivos militares e era uma das formas das forças armadas norte-americanas manterem as comunicações em caso de ataques inimigos que destruíssem os meios convencionais de telecomunicações.

Em 1965, criou-se o primeiro computador de rede do mundo um computador TX-2 em *Massachussets* com um Q-32 na Califórnia com uma linha discada de baixa velocidade. O resultado deste experimento foi a comprovação de que computadores poderiam trabalhar bem juntos, rodando programas e recuperando dados quando necessário em máquinas remotas, mas que o circuito do sistema telefônico era totalmente inadequado para o intento. Foi confirmada assim a convicção sobre a necessidade de trocas de pacotes.

O DoD, em 1969, promoveu a criação de um sistema de comunicações que permitiu a interligação dos principais computadores da ARPA dando origem assim a ARPANET, conectada a quatro servidores: a Universidade da Califórnia, em *Los Angeles* e em Santa Bárbara, o Instituto de Pesquisa de *Stanford* e a Universidade de *Ytah*, em *Salt Lake City.*

Uma rede cooperativa, chamada de *Bitnet* (*Because It's Time NETwork*), em 1981, inicia na *City University*, de *Nova York*, oferecendo correio eletrônico, servidores de lista e transferência de arquivos. A *Bitnet* se torna uma alternativa à Internet.

Na década de 1980 a *Arpanet* adota um protocolo que permite a qualquer tipo de computador se conectar à rede, o TCP/IP (*Transmission Control Protocol/Internet Protocol*). O nome *Internet* começa a ser utilizado para definir as redes que estão conectadas entre si

através deste protocolo. A nova forma de conexão permite que pequenas instituições sejam ligadas à rede.

Em 1992, a *internet* já conectava um milhão de computadores e passou a ser utilizada também para fins comerciais. Foram então criados o *Archie* (um sistema de busca em arquivos) e o GOPHER (um sistema de organização da informação na Internet na forma de menus e bancos de dados).

Foi somente no ano de 1990 que a *internet* começou a alcançar a população em geral. Neste ano, Tim Berners-Lee teve a idéia de desenvolver com sua equipe do CERN (*European Organization for Nuclear Research*, de Genebra), um sistema de hipertexto que deveria funcionar em redes de computadores, a linguagem baseada em SGML(*Standard Generalized Makup Language*) que ele denominou de HTML (*HiperTxt Markup Language*).

Nesse momento, ele pensava apenas nos cientistas que precisavam compartilhar suas pesquisas uns com os outros. Esses pesquisadores, em 1991 tiveram a idéia de criar a *World Wide Web*. No início a maior parte das informações ainda era no formato de texto, com poucos desenhos. Em 1992, *Marc Andressen*, do NCSA (*National Center for Supercomputer Activity*), criou o primeiro navegador para *Internet*: o *Mosaic*, para sistema do Windows. Em seguida apareceram versões do *Mosaic* para *Macintosh* e *Microsoft Windows*. O *Mosaic* era capaz de interpretar gráficos e realizar navegações através de *links*, como podemos ver atualmente na *Web*.

Neste período a *internet* permitia ao usuário somente trocar mensagens por via eletrônica ou transferir dados utilizando protocolos comuns de comunicação. Foi então que em 1989 que um físico inglês, Tim Berners-Lee, começou a trabalhar num sistema que deveria ser capaz de gerenciar documentos de todos os tipos e formatos como gráficos, desenhos, relatórios, etc. e que utilizaria *links* hipertextuais para interligar as páginas de uma forma não onerosa. Começou a surgir então *a WWW- World Wide Web* um desdobramento da *internet* que permite a difusão e transferência de informações e arquivos multimídia através de *hiperlinks* .

Em 1994, Berners-Lee fundou o consórcio W3C uma organização destinada a desenvolver padrões e tecnologias que possa beneficiar a sociedade através da oferta de novas formas de comunicação entre humanos e oportunidades de compartilhamento de conhecimento.

> A missão do W3C é levar a *web* ao seu potencial máximo, através do desenvolvimento de tecnologias (especificações, diretrizes, *software* e ferramentas) e criar um fórum para informação, comércio, inspiração, pensamento independente e compreensão coletiva.(BERNERS-LEE, 2002)

Atualmente o W3C tem mais de quatrocentos e cinqüenta membros e um quadro de aproximadamente setenta pessoas a nível global dedicadas em tempo integral que contribuem para o desenvolvimento de especificações de W3C e *software*.(Dados do Site do W3C)

7.2 A World Wide Web- WWW

A *World Wide Web* ou *Web*, de acordo com Teixeira (1997) é uma coleção de documentos hipertextos ligados entre, criando um mundo de informações digitais que envolvem texto, imagens e sons, construindo-se em um dos maiores acervos multimídia que integra as tecnologias de comunicação, transmissão de imagens e sons, criando uma verdadeira rede de difusão de conhecimento.

Esses documentos são as chamadas "páginas", que são arquivos de computador com variados tamanhos (número de caracteres) e apresentam as seguintes características:

- um endereçamento conhecido como *Unifom Resourse Locator* (URL) que localiza o arquivo num computador ligado à rede;
- um protocolo de tranferência, o *Hypertext Transfer Protocol* (http) que faz a interligação entre o computador do usuário e o local onde a página está localizada (servidor ou *host*);
- uma linguagem de marcação padrão que estrutura e define os componentes das páginas na web, como a *Hipertext Markup Language* (HTML).
- utiliza um programa navegador como o Internet Explore que percorre uma rede de documentos vinculados, interpreta a linguagem HTML e exibe-a na tela do computador;

Os URLs contêm várias partes. Por exemplo, num endereço como: http://www.unb.br, a primeira parte – a http://- detalha qual protocolo da Internet usar. A segunda – a parte que geralmente tem um "www" -, normalmente informa que tipo de recurso *internet* está sendo conectado. A terceira parte – "unb" pode variar em comprimento, e identifica o servidor da rede a ser conectado. A parte final identifica um diretório específico no servidor e uma *home page*, documento e ou outro objeto da *internet*.

A maioria das páginas estão escritas em linguagem HTML. Essa linguagem é uma evolução da linguagem SGML- *Standard Generalized Markup Language*, uma linguagem padronizada que utilizou pela primeira vez as "marcas" ou "tags", um conjunto de códigos pré definidos que definem componentes relacionados com a aparência e a funcionalidade das página além de indicar o início e o fim da estrutura de compõe o documento. A linguagem HTML também é composta por um número fixo de *tags* que definem a aparência da página.

A linguagem HTML é muito simples e pode ser criada utilizando-se qualquer editor de texto. Sua simplicidade não a limita, pois a mesma consegue utilizar uma grande quantidade de recursos como a utilização de *frames* (janelas), e outros recursos multimídia.

Uma página HTML pode conter *tags* que especificam URLs de outra páginas, constituindo assim os conhecidos *links*. Estes *links* utilizam termos de indexação não controlados, atribuídos pelas pessoas que criam os *sites* e em geral elas não fazem nenhum controle sobre os termos que serão utilizados. Eles não estão fisicamente armazenados, de forma que não é possível, por exemplo, determinar quais são as páginas que referenciam uma página específica.

Veja abaixo um exemplo de linguagem HTML:

```
<html>
  <head>
  <title>Exemplo de HTML</title>
  </head>
  <body>
  <h1>Exemplo de HTML</h1>
  bla bla bla
  </body>
        </html>
```

O significado das *tags* são facilmente decifráveis. Cada *tag* HTML, ou instrução fica entre um sinal de menor que e um sinal de maior que: <p>. Os marcadores **<html>** e **</html>**

delimitam a descrição da página que é dividida em duas partes: o preâmbulo e o corpo da página.

O preâmbulo é delimitado pelos marcadores **<head>** e **</head>**. O preâmbulo pode conter, entre outros, o título do documento que vai ser apresentado no topo da janela do paginador e meta-informações ou metadados (isto é, informações que descrevem de alguma forma o documento como palavras chave, resumo, etc.).

O preâmbulo da página utilizada como exemplo para ilustrar a anatomia de página WWW contém apenas um título. Tal título é delimitado pelos marcadores **<title>** e **</title>**.

O corpo de uma página é delimitado pelos marcadores **<body>** e **</body>** e contém as informações a serem apresentadas na área de visualização do paginador. No corpo do exemplo temos um título em nível 1 (existem 6 níveis sendo o nível 1 o mais alto) delimitado pelos marcadores **<h1>** e **</h1>** e um texto qualquer.

8 Sistemas de Busca da Web

O surgimento da *internet* trouxe consigo o problema de recuperar informações devido a grande explosão das publicações disponibilizadas por meio delas. Para tentar amenizar o problema, criou-se desde o início, ferramentas utilizadas para a localização de recursos informacionais como o *Archie*, criado em 1990, composto de um banco de dados com nomes de arquivos da *web* que buscava arquivos em repositórios de FTP; o *Gopher* que recupera informações mediante sistemas de *menus* hierárquicos permitindo a recuperação de todo tipo de informação textual, e o *Veronica* que utiliza palavras-chave para localizar informações em servidores *Gopher*.

Os sistemas de busca de acordo com Lopes (2006), surgem a partir de 1994, inicialmente oriundos das atividades de pesquisa e de outros profissionais atuantes na *web*, sendo que o ponto de referência conhecido era a *World Wide Web Virtual Library*, no *site* do CERN, consistindo numa lista alfabética de assuntos com *links* de páginas, instrumento de auxílio que atualmente é classificado como ferramenta de busco do tipo diretório.

De acordo com Yamaoka (2002) existem duas abordagens básicas de sistemas de busca na *web*: os diretórios e os mecanismos de busca (*search engine*). Para o autor, um

sistema de busca pode manter e oferecer simultaneamente os serviços de diretório e mecanismos de busca.

8.1 Diretórios

De acordo com Cendón (2001), os diretórios foram à primeira tentativa de se solucionar o problema da recuperação de informação na *web* e precedeu os mecanismos de busca por palavras-chave, numa época em que o conteúdo da *web* ainda podia ser coletado e indexado de forma manual.

Para Yamaoka (2002) os diretórios são listas de assuntos organizadas em categorias, geralmente com uma estrutura hierárquica (árvore).

Alonso (2004) complementa a afirmação de Yamaoka e diz que os diretórios são guias ou listas agrupadas e ordenadas sistematicamente por categorias e subcategorias, que registram as direções e uma pequena descrição dos diferentes sites ou recursos disponíveis na *internet*, (..). Estes são definidos manualmente por uma equipe especializada do diretório, portanto sua atualização não é automática.

O uso dos diretórios são mais apropriados para buscas sobre temas mais amplos e de pouco domínio do usuário. Alguns mecanismos de busca possuem diretórios também, como é o caso do *Google*.

Os diretórios, segundo Feitosa (2006), surgiram com a intenção de se coletar manualmente, ou por meio de indicações de usuários, a maior quantidade de informação possível, contando-se a grande variedade dos assuntos disponíveis na Internet.

Segundo Yamaoka (2002), os diretórios buscam como abordagem principal:

- Manutenção do nível de qualidade estabelecido;
- Classificação precisa de sites na *web*.

Nos diretórios o autor de uma página *web* cadastra a URL de sua página associando a ela uma ou mais categorias, as quais podem conter subcategorias, que descrevem o assunto tratado na página. Neste momento os sites recebem uma classificação hierárquica de assunto e permitem ao usuário localizar as informações também nas subcategorias. Geralmente nos

diretórios são incluídas outras áreas de interesse mais amplo para chamar atenção do usuário, como: educação, esporte, viagens e outros.

Uma outra característica dos diretórios é que cada categoria de assunto é também uma página da *web*. A página de uma determinada categoria é formada por um conjunto de *links* para as páginas relacionadas àquela categoria e um conjunto de *links* para a sub-categoria.

O método utilizado pelos diretórios possui inevitáveis desvantagens, mas também enormes vantagens. A cobertura temática nem sempre é completa e regular, se o usuário deseja fazer uma busca de um determinado assunto que não se enquadra dentro de nenhuma categoria pré-esbelecida, ou se o assunto é uma combinação de categorias, o resultado obtido pode não ter a precisão esperada. Além disso, a recuperação da informação nesses sistemas de busca, geralmente exige do usuário uma identificação preliminar da área em que o tema de interesse pode estar armazenado, ocasionando sempre um tempo maior de busca, segundo observa Lopes (2006).

Por outro lado, se a busca do usuário está relacionada diretamente com as categorias existentes, é possível que ele obtenha alta precisão.Geralmente as páginas indexadas pelos diretórios possuem *links* para outras páginas de assuntos correspondentes.

Uma outra questão é que os sites coletados pelos diretórios passam por um processo de seleção humano, por meio de sugestões ou até mesmo pesquisas na *web*. O problema é que centenas de sites podem ser acrescentados semanalmente, e quantidade nem sempre é sinônimo de qualidade.

8.1.1 Tipos de diretórios

Segundo Cendón (2001), os diretórios embora tenham características genéricas, variam quantos aos princípios de organização, à forma de descrição dos *sites* e os assuntos cobertos, apresentando características próprias.

Quanto aos princípios de organização podem utilizar relações hierárquicas ou os esquemas tradicionais de classificação, como o sistema de cabeçalhos de assuntos. São geralmente mantidos por profissionais da informação ou bibliotecários.

Já a descrição de *sites* pode limitar-se a incluir títulos e breves resumos de até trinta palavras, ou fornecer descrições criteriosas e detalhadas dos recursos, podendo incluir críticas

e até mesmo avaliações dos mesmos. Os últimos são geralmente chamados de diretórios avaliativos ou acadêmicos.

Quanto ao assunto de cobertura os diretórios podem ser temáticos, quando abrangem diversas áreas de conhecimento, ou especializados, quando se dirigem a um tipo específico de usuário e, portanto abrangem áreas específicas de cada assunto.

Os diretórios mais conhecidos segundo informações do *site Search Engine Showdown* e outras fontes consultadas são: *Yahoo* Brasil (http://www.yahoo.com.br); *LookSmart* (http://search.looksmart.com/) *Britannica* (http://britannica.com); *The Open Directory-DMOZ* (http://dmoz.org/). Abaixo serão melhor descritos:

O **Yahoo** é o diretório mais popular da *web* segundo Feitosa (2001), foi criado em 1994 por dois estudantes de engenharia elétrica, David Filo e Jerry Yang, que estavam interessados em organizar uma coleção de seus sites prediletos. A quantidade de páginas referenciadas cresceu rapidamente e logo os estudantes foram obrigados a reorganizá-lo para tornar-se um diretório de busca local. O *Yahoo* possui uma base de dados bem grande devido ao seu tempo de existência, muitos serviços e produtos para informação geral e popular. Tem como desvantagem de uso a grande ênfase comercial.

Veja na página seguinte sua interface:

Figura 1: Interface do Diretório *Yahoo* na web.

O **LookSmart** também é bastante conhecido, possui comandos que possibilita que se refine a pesquisa e ainda oferece tópicos relacionados com o termo de busca empregado pelo usuário, mas seu diferencial está em possibilitar a organização, a busca e o compartilhamento de páginas da *web* por meio da criação de uma "*web* pessoal" na qual o usuário salva as páginas encontradas na *web* possibilitando que as encontre novamente no instante de seu acesso.

Esse serviço possibilita ainda que o usuário indexe a página, separe-a por área, avalie o *site* e envie um *link* da página para *e-mails*. Outra grande vantagem é que ele possui um banco de dados com mais de dez milhões de artigos, divididos por área. Suas desvantagens é que não possui pesquisa avançada na página principal do diretório, mas somente no banco de dados dos artigos e muitos desses artigos são pagos.

Figura 2: Interface do Diretório LookSmart na *web*.

O **Britannica**, anteriormente conhecido como BLAST (*Encyclopedia Britannica Links and Search Tool*) e BIG (*Encyclopædia Britannica's Internet Guide*), a partir de novembro 1999, passou uma importante mudança, crescendo e transformando-se num grande diretório da *web*. Nesse período passou a incluir o conteúdo completo da *Encyclopedia Britannica* e

integrar diversos periódicos. Se caracteriza por dar acesso a artigos de texto integral de aproximadamente setenta periódicos. Seu diferencial está em apresentar grande foco acadêmico, além de possuir um dicionário escolar e um tesauro. Apresenta como desvantagem o fato do diretório não ser separado da pesquisa na *web*.

Figura 3: Interface do Diretório Britannica na *web*.

O **Open Directory Project**, anteriormente conhecido como *NewHoo* e agora como DMoz é um dos mais amplo e abrangente diretório da *web* editado por humanos. Ele é construído e mantido por uma vasta comunidade global de editores voluntários.O *Open Directory* foi fundado dentro do espírito do *Open Source Movement* (Movimento pelo software livre e com códigos fontes abertos), e é o único grande diretório que é totalmente livre e gratuito. Não há nenhum custo para a submissão de sites ao diretório, e para o uso de seus dados.

Os dados do *Open Directory* são disponibilizados gratuitamente a qualquer um que concorde com os termos de nossa licença de uso. Apresenta como vantagem a possibilidade de acesso em língua portuguesa, uso da arquitetura RDF, desenvolvida pelo consórcio W3C e

o acesso aberto a qualquer usuário. Sua desvantagem é que por possibilitar que qualquer possa ser editor, sua qualidade pode se tornar inconsistente.

Figura 4: Interface do Diretório DMOZ na *web*.

8.2 Mecanismos de Busca (Search engine)

Os mecanismos de busca também chamados de *search engines*, *sites* de busca ou portais, são mecanismos que permitem ao usuário submeter sua expressão de busca e recuperar um lista de endereços de páginas (URLs) que satisfaçam sua necessidade de informação. Começam a surgir quando a *web* tornou-se bastante complexa de forma que era impossível indexar manualmente todas as páginas da *web*. Foram inicialmente criados por estudantes de pós-graduação, professores, analistas de sistemas e outras pessoas interessadas em recuperar os documentos da *web*.

De acordo com Cendón (2001), o ALIWEB (*Archie-Like Indexing on the Web*) e Harvest são exemplos das primeiras tentativas de criar mecanismos de busca por palavras-

chaves, e utilizavam tecnologias diferentes das atuais. Logo após, houve a criação do *Archie* em 1990, um buscador de arquivos em repositórios de FTP; o *Gopher* um buscador de informação textual, e o *Verônica* um buscador que localiza informações em servidores *Gopher*.

Em 1994, surge o primeiro mecanismo de busca baseado em robôs o *WebCrawler*. Antes dele, de acordo com Feitosa (2006) um usuário podia pesquisar apenas nas URL's ou em descrições de páginas fornecidas pelos seus autores. O *WebCrawler* tornou- se bastante popular e devido a grande quantidade de acesso.

Os mecanismos de busca utilizam três componentes: um programa de computador, uma base de dados, também conhecida como índice ou catálogo e um programa de busca. Esse programa é acionado pelo usuário ao realizar uma pesquisa na *web* com seus termos de busca e as respostas são apresentadas a partir dos dados e endereços contidos na base de dados do mecanismo. (LOPES, 2006, p. 21).

Numa visão mais abrangente Yamaoka (2002) cita que os mecanismos de busca apresentam três funções principais: um robô, que localiza os documentos; um indexador, que extrai as informações dos documentos e uma interface com o usuário.

Os robôs são "programas que o computador hospedeiro da ferramenta de busca lança regularmente na Internet, na tentativa de obter dados sobre o maior número possível de documentos para integrá-los, posteriormente, à sua base de dados" (Cendón, 2001, p. 41).

Esses robôs "viajam" através da *web* a fim de selecionar URLs de páginas de potencial interesse para quem deseja indexá-las. Utilizando a metáfora da *internet* como "Teia mundial" os robôs são também chamados de *spiders* (aranha) ou ainda robôs, *crawlers* ou *worms* que rastreiam a "Teia".

Segundo Robredo (2006), o fundamento do funcionamento dos mecanismos de busca baseiam-se nos seguintes princípios, listados abaixo:

- Armazenam informações sobre grandes quantidades de páginas na *web* recuperadas na rede, analisam o conteúdo, indexam as páginas e as armazenam em bancos de dados;
- As palavras-chave utilizadas pelos usuários em suas perguntas são comparadas com as entradas das bases de dados indexadas, para selecionar as páginas pertinentes;

- Podem ordenar as páginas recuperadas segundo critérios de maior e menor relevância que variam de um mecanismo de busca a outro.

A indexação automática dos mecanismos de busca é feita inicialmente por meio de seleção de endereços (URLs) de páginas. Nessa fase os robôs rastreiam a estrutura hipertextual da *web* colhendo informação sobre as páginas que encontram. Para reduzir os impactos da estrutura complexa da *web* os robôs podem utilizar duas estratégias: a primeira chamada de *breadth-first* que faz uma busca mais superficial pelos níveis de *site*, aumentando a amplitude da pesquisa e a segunda que faz uma busca em *links* de um mesmo servidor aumentando assim o maior detalhamento de um assunto tratado no *site*, chamada de *deep-first*.

Após recolhidas as URLs dos *sites* os robôs as adicionam a sua base de dados. Para aumentar a velocidade de cobertura da *web* podem ser usados vários robôs trabalhando em paralelo, cada um cobrindo uma região ou domínio diferente da *web* e enviando as informações para a base de dados.

A próxima etapa após a criação do banco de dados composto pelas URLs é encaminhar os documentos aos indexadores, que extraem as informações das páginas HTML e a armazenam em suas base de dados. No processo de indexação automática, um algoritmo (conjunto de operações elementares, organizadas logicamente) realiza, em certa medida, o trabalho do indexador no processo de escolha dos termos significativos. (ROBREDO, 2005, P. 170). Os termos mais significativos (descritores) são retirados do título, ou do próprio texto e resumo, criando assim, os índices, chamados em linguagem técnica de arquivos invertido, que são utilizados para dinamizar a busca de informações na sua base de dados.

8.2.2 Tipos de mecanismos de busca

Os mecanismos de busca em geral possuem as características acima descritas, porém, a maior parte deles apresentam perfis que lhes são próprios. Eles diferem quanto:

- O tamanho dos bancos de dados;
- Os critérios para a indexação;
- Os critérios para a inclusão de páginas;
- A freqüência de atualização de dados e ordenação de resultados.

O tamanho das bancos de dados é medido de acordo com a quantidade de URLs. Esse tamanho é um dos indicadores de qualidade do mecanismo de busca, pois uma página só pode ser encontrada se algum mecanismo de busca a tiver incluído. No entanto, é impossível que algum mecanismo de busca consiga incluir todas as páginas existentes na *web*. Esse tamanho é um dos fatores de limitação da recuperação da informação por parte dos sistemas de busca, já que os robôs não conseguem indexar todas as páginas existentes na *web*, além disso, existe uma chamada "*Web Oculta ou invisível*" que segundo a literatura pesquisada é muito maior que a *web* que se conhece. A *Web Oculta* será melhor descrita mais adiante.

A respeito critérios de indexação a maioria dos mecanismos de busca indexam, palavras do texto visível nas páginas. Porém, algumas podem utilizar *tags* para restringir ou ponderar a indexação das páginas a determinadas marcas localizadas nas páginas, como a utilização de termos incluídos nos *metags* de classificação, de descrição e de palavras-chave ou até textos associados a imagens. Os *metatags* de classificação fornecem palavras-chave que define o conteúdo da página. Os *metatags* de descrição recuperam a descrição da página feita pelo próprio autor. E os *metatags* de palavras-chave recuperam as próprias palavras-chave designadas pelo próprio autor no momento de sua criação.

Quanto aos critérios de inclusão, é importante ressaltar que a maioria das URLs são salvas, mas apenas algumas páginas são indexadas por causa da política de indexação das empresas. Alguns mecanismos de busca incluem apenas a *home page* e algumas páginas principais.Um problema que pode ocorrer é a duplicidade de URLs na base de dados o que diretamente sua qualidade. Esse problema pode ser reduzido através da utilização de

algoritmos que são capazes de detectar semelhanças entre as páginas da *web* em diversos formatos.

A freqüência de atualização de dados é a medida com que os robôs revisam periodicamente a *internet*, não só para incluir novas páginas, mas também para deletá-las ou incluir as modificações das que já existem no índice. Caso os robôs não façam essa revisão, páginas não mais existentes ou com informações diferentes podem ser recuperadas sem serem relevantes. Cada robô possui sua própria política de atualização.

A ordenação de resultados é feita com a finalidade de permitir que os sites de maior relevância apareçam em primeiro lugar.Assim, a maioria utiliza algoritmos de ordenação baseados em critérios como: a localização e freqüência de ocorrência das palavras em uma página, a densidade que é o número de termos da consulta que estão presentes na página e a proximidade dos termos e o número total de vezes que uma palavra ocorre num banco de dados.

Segundo a literatura pesquisada, os mecanismos de busca mais conhecidos e citados de acordo com o *site Search Engine Showdown* são: *Google* (www.google.com.br); *Ask* (www.ask.com); *Live Search* (www.live.com) e *Yahoo* (http://br.search.yahoo.com/). Esses mecanismos serão melhor descritos abaixo:

O **Google** atualmente é o mecanismo de busca mais conhecido e usado. Era originalmente um projeto da universidade de *Stanford* dos estudantes Larry e Sergey Brin chamado de BackRub. Por volta de 1998, o nome mudou para *Google*, e o projeto saltou fora do campus e acabou ganhando o mundo. *Google* fornece a opção para encontrar mais do que *Web pages*, no alto da caixa da busca no *Home Page* de *Google*, pode-se facilmente procurar imagens através da *web*. Igual a outros mecanismos de busca, as temáticas aparecem na página principal. Apresenta a maior base de dados existente e se atualiza com grande regularidade. A busca no *Google* só devolve páginas que incluem os termos introduzidos pelo usuário.

Outra característica que o diferencia dos demais mecanismos de busca, é que ele analiza a proximidade das expressões empregadas para a busca nas páginas. Essas páginas possuem prioridade no momento demostrar os resultados. No lugar de resumo de páginas, este mecanismo de busca mostra o texto coincidente do documento recuperado com as palabras chaves solicitadas nos termos de busca em negrito e o total de páginas encontradas.

Utiliza o software *PangeRank*(TM), um sistema para dar notas em páginas na *web*, desenvolvido pelos fundadores Larry Page e Sergey Brin na Universidade de Stanford este *software* considera um *link* em cada página como um voto de forma a disponibilizar os mais

acessados em primeiro lugar nos resultados das busca. Sua interface é demonstrada na página seguinte:

Figura 5: Interface do Google na *web*.

O **Ask** foi lançado em 1997 e era anteriormente conhecido como *Ask Jeeves*, passou por várias mudanças até se transformar num mecanismo de busca da *web*, usando uma base de dados originalmente desenvolvida pelo *Teoma*. A tradição de pergunta-resposta continuou no *Ask* e o sistema por meio desse método oferece resposta direta a perguntas em linguagem natural. *Ask Jeeves* trocou a base de dados do *Teoma* pelo *Direct Hit* em janeiro de 2002, possui atualmente uma base de notícia, imagens e *blogs*.

Apresenta como vantagem buscas em comunidades da *web*, inspeção prévia dos *sites* recuperados nas buscas, conversões de medidas, busca por pessoas famosas e principalmente

a busca por perguntas em linguagem natural. Como desvantagem o mecanismo não possui uma base de dados tão grande e alguns sites são pagos.

Figura 6: Interface do Ask.

O **Live Search** é o sucessor do MSN *Search.* Às vezes é chamado somente Live.com ou Windows *Live Search*, isto é o mecanismo de busca da Microsoft na *web*. Lançado em setembro de 2006, ele usa um banco de dados próprio e único que possibilita a busca específica por notícias, imagens, vídeo, alimentos, pesquisa acadêmica entre outros. Apresenta como principais vantagens o tamanho de seu banco de dados que é um dos maiores, boa estrutura de pesquisa avançada, busca booleana, opções de busca por locais

específicos. Como desvantagem não oferece truncamento e a pesquisa avançada não está na página inicial.

Figura 7: Interface Live Search.

O **Yahoo** é um sistema híbrido de forma que também pode ser considerado como um mecanismo de busca. Possui diversas bases de dados como: imagens, vídeos, *shopping*, notícias, áudio e outras. È um dos mecanismos que devido ao tempo de existência apresenta como vantagens uma grande base de dados, além de *links* para o Diretório *Yahoo*, busca booleana, busca por proximidade em palavras e frases. Tem como principais desvantagens o uso de truncamento somente na busca avançada, a exigência de *links* em https:// e pagamento para inclusão de *sites*.

Figura 8: Interface do Yahoo Search.

Atualmente o Yahoo desenvolveu um mecanismo de busca que possibilita ao usuário fazer perguntas e obter respostas reais de pessoas reais: o Yahoo Respostas (http://answers.yahoo.com). É bastante divertido e interativo porque além de fazer perguntas sobre qualquer assunto o usuário pode ajudar outras pessoas respondendo perguntas. Diferente do correio eletrônico e das salas de bate papo, as perguntas do Yahoo Respostas são resolvidas quando a melhor resposta é escolhida, podendo a comunidade então avaliá-la. Quando o usuário faz a pergunta, perguntas similares aparecem e se não houver nenhuma pergunta que satisfaça sua necessidade ele pode optar em receber um *e-mail* notificando-o toda vez que alguém responder sua pergunta.

Figura 9: Interface do Yahoo Respostas.

Em relação à utilização dos mecanismos de busca, a figura abaixo demonstra que o *Google* e o *Yahoo* continuam liderando as pesquisas de busca executadas na *web* pelos internautas americanos:

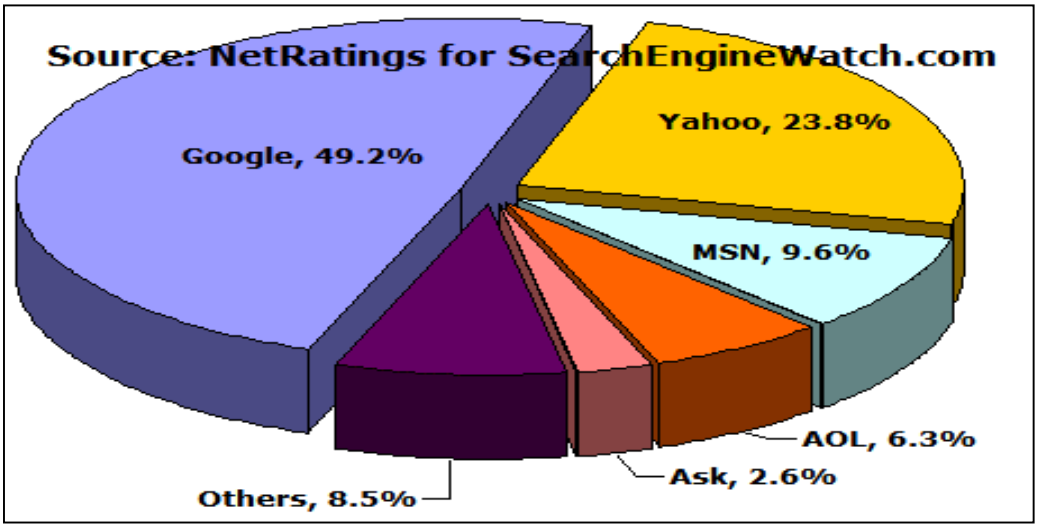

Figura 10: Utilização dos serviços de busca na *web*, por internautas americanos em novembro de 2006.

Fonte: (Search Engine Watch, 2006)

8.3 Metabuscadores

Os metabuscadores são sistemas que permitem a execução de uma mesma busca em mais de uma ferramenta (mecanismos ou diretórios), podendo exibir numa só lista ao mesmo tempo todos os resultados encontrados nos outros sistemas de busca.

Segundo Lopes (2006) esses sistemas não possuem nenhuma base de dados própria e sim um *software*, que pesquisa base os dados solicitados por um usuário nos outros mecanismos de busca, apresentando os resultados num formato em que a quantidade de respostas é fornecida para cada mecanismo de busca em separado.

Assim como os mecanismos de busca, os metabuscadores também possuem algumas diferenças entre si:

- Quanto à interface: muitos fazem as buscas entre 6 ou 10 mecanismos maiores, outros oferecem a opção de escolha sobre em quais mecanismos de buscaes pesquisar ou até mesmo a opção sobre que ferramenta de idiomas utilizar.
- Quanto ao processamento de consultas: alguns possuem a formulação de expressão de buscas livre ou utilizam operadores booleanos (AND, OR, NOT). E quanto ao tempo de resposta as consultas nas ferramentas de busca podem ser feitas de forma seqüencial ou simultaneamente.

Hock (2001 apud LANCASTER, 2004, p. 344) acrescenta outras diferenças entre os metabuscadores:

- Os mecanismos específicos que abrangem;
- A capacidade de repassar consultas mais complexas – como as que incluem expressões, enunciados booleanos, etc. - para os mecanismos de busca ‘alvo’;
- A forma como a saída é apresentada, inclusive se eliminaram ou não registros certos encontrados em duplicata em vários mecanismos.

Os metabuscadores são indicados para fazer pesquisas nas quais são encontrados poucos resultados, verificar quais os mecanismos de buscas individuais trazem as melhores respostas e fornecer uma visão mais ampla do resultado de cada ferramenta.

Porém o seu uso também traz algumas desvantagens, pois os mesmos não possuem os mecanismos de refinamento das pesquisas utilizados pelos outros mecanismos de busca e como conseqüência, obtém-se alta revocação de resultados e baixa precisão.

Rock (2001 apud LANCASTER, 2004, p. 344) aponta que os três principais pontos fracos dos metabuscadores são:

1) muitas vezes limitam estritamente o número de registros que recuperarão de cada mecanismo (às vezes não mais de dez)
2) muitas vezes não repassam aos mecanismos consultas que tenham um mínimo de complexidade;
3) na maioria dos casos, só fazem buscas em dois ou três dos maiores mecanismos de busca [...]

Braski (2004) resume as idéias de Hock (2001) quando considera que as deficiências dos metabuscadores estão basicamente relacionadas com a forma de apresentação dos resultados em uma só interface e com a incapacidade de manipularem pesquisas complexas.

Alguns dos metabuscadores mais conhecidos de acordo com a literatura consultada são: Vivíssimo (www.vivissimo.com); *Clusty* (www.clusty.com); *Ixquick* (www.ixquick.com); *Dogpile* (www.dogpile.com); *Mamma* (www.mamma.com); *Kartoo* (www.kartoo.com); *Metacrawler* (www.metacrawler.com).

O **Vivísimo** foi fundado originalmente em 2000 por três cientistas Universitários que decidiram resolver o problema da sobrecarrega de informação em procura da *web* por meio do de uma nova tecnologia. Usaram um algoritmo matemático e conhecimento lingüístico profundo para fazer relacionamentos entre termos de procura. Ao decorrer dos anos, a companhia construiu uma tecnologia original de *clustering*.

O Mecanismo *Clustering* é uma parte integrante da plataforma de busca do Vivíssimo, que com ajuda de seus operadores localizam a informação que os usuários necessitam num contexto específico. Em vez de retornar somente os primeiros dez resultados de milhares sobre milhares de tentativas, os resultados da busca estão agrupados junto por categorias.

Estas categorias são criadas dinamicamente no momento da busca.

Figura 11: Interface do Vivíssimo na *web*.

O **Clusty** foi fundado em 2004 pelo Vivíssimo é um metamecanismo que faz a busca em vários mecanismos de busca, combina os resultados e gera uma lista baseada em classificação por comparação entre os primeiros resultados selecionados pelos mecanismos de busca. Este método de aproximação, "metasearch", seleciona os melhores resultados ao topo da lista recuperada e para o *spam* de mecanismo de busca. Mas o que realmente torna o *Clusty* raro é que da mesma forma que o Vivíssimo, ao invés de apresentar vários resultados da busca em uma longa lista, o mecanismo de busca agrupa resultados semelhantes em categorias por meio do chamado "Mecanismo Clusty", também usado pelo Vivíssimo.

Figura 12: Interface do Clusty na *web*.

O **Ixquick** assim como os outros metabuscadores, também faz a busca em diversos outros mecanismo, possui um método de classificação sobre os resultados da busca, no qual ele concede uma estrela (★) para cada resultado escolhido como um dos dez melhores. Oferece como estratégia de busca, a lógica booleana, aproximação por frases, e buscas por campo, seu diferencial está no fato buscar em mais de dezoito línguas diferentes, entre elas o chinês e o coreano. O Ixquiz oferece também a possibilidade de marcações sobre resultados da busca, (✓) , para documentos relevantes e (✗) para os irrelevantes, economizando assim o tempo para o usuário, já que ele não precisará voltar em resultados já excluídos. Dessa forma se um resultado é marcado positivamente, resultados semelhantes são trazidos na próxima busca. Outra vantagem é que o *Ixquick* oferece também versão em língua portuguesa.

Figura 13: Interface do Ixquiz na *web*.

O **Dogpile** foi construído para buscar os melhores resultados disponíveis na *web*. Isto é conseguido através da busca nos mecanismos mais populares e por recobrar os melhores resultados combinados. Os fundadores desse metabuscador correlacionaram a busca na *web* à ajuda de um cão de caça virtual chamado de Arfie que é a marca do *Dogpile*. Uma vez recobrados os resultados, a tecnologia inovadora de *metasearch* usado pelo *Dogpile* trabalha, retirando duplicatas e analisando os melhores resultados.

Figura 14: Interface do DogPile na *web*.

Criado em 1996 por uma tese de mestrado, o **Mamma** ajudou a introduzir a idéia de *metasearch* à *internet*. Devido a seus resultados de qualidade e os benefícios de *metasearch*, o Mamma cresceu rapidamente entre conversas informais e rapidamente tornou-se um dos maiores metabuscadores da *internet*. A capacidade do Mamma em reunir os melhores resultados disponíveis nos maiores mecanismos de busca da *web* e fornecer ferramentas úteis a seus usuários resultaram em Prêmios na categoria de "Melhor Metasearch" do concurso anual do *Search Engine Watch.*

Figura 15: Interface do Mamma na *web*.

O **Kartoo** usa mapas interativos para a apresentação de resultados. Logo que uma busca é lançada, o Kartoo analisa a pergunta, interroga os mecanismos mais relevantes, seleciona os melhores locais e os coloca num mapa. Os *sites* encontrados são representados de acordo com a sua relevância pelo tamanho do desenho dos mapas. Quando se move pelos desenhos, as palavra-chaves são iluminadas e uma descrição breve do local aparece ao lado esquerdo da tela, permitindo o refinamento da busca.

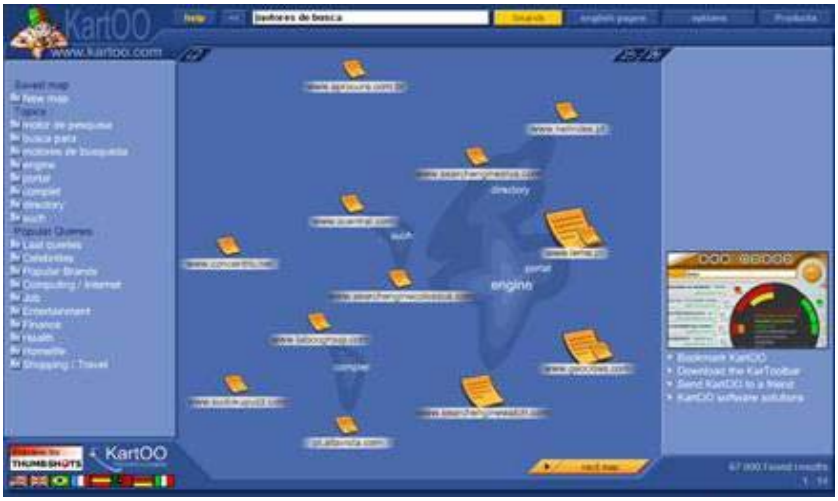

Figura 16: Interface do Kartoo na web.

O **MetaCrawler** foi originalmente desenvolvido em 1994 na Universidade de Washington. O metabuscador uniu-se a Rede de InfoSpace em 2000 por quem é operado atualmente. O *MetaCrawler* também faz a busca em vários mecanismos de busca da *web* incluindo *Google*, *Yahoo*, *MSN*, *ASK*, *LookSmart* e outros. Sua vantagem é que seus resultados são apresentados em uma única interface.

Figura 17: Interface do Metacrawler na *web*.

O quadro abaixo apresenta as características gerais dos principais metabuscadores apresentados:

Metabuscador	Proprietário	Bases de dados que abragem	Bases de dados adicionais	Características especiais
Vivisimo	Vivisimo	Ask, MSN, Gigablast, Looksmart, Open Directory, Wisenut	Google	Resultados por categorias
Clusty	Vivisimo	Ask, MSN, Gigablast, Looksmart, Open Directory, Wisenut	Google	Resultados por categorias
Ixquick		AltaVista, EntireWeb, Gigablast, Go, Looksmart,Netscape, Open Directory,Wisenut, Yahoo	Yahoo	Marcações dos resultados
Dogpile	InfoSpace	Ask, Google, MSN, Yahoo!, Teoma, Open Directory e outros	Google, Yahoo	Abrange os 4 maiores mecanismos
Mamma	Mamma	Ask, Google, MSN, Yahoo!, Teoma, Open Directory, more	Miva, Ask	Opções mais refinadas
Kartoo		AlltheWeb, AltaVista, EntireWeb, Exalead, Hotbot, Looksmart, Lycos, MSN, Open Directory, Teoma, ToileQuebec, Voila, Wisenut, Yahoo		Exibe resultados visuais

Tabela 1: Visão geral dos metabuscadores
Fonte: Search Engine Showdown (2006)

Metabuscadores extintos:

- C4 (anteriormente *Cyber* 411), extinto desde dezembro 2002.
- *Inference Find*, extinto desde março 2001.
- *MetaFind* fundiu-se ao *MetaCrawler* por volta de janeiro do ano 2000.
- *SavySearch*, comprado pelo Search.com em 1999.

9 Estratégias de Busca

De acordo com Rowley (2004), estratégia de busca é o conjunto de decisões e ações tomadas durante uma busca. A autora completa sua afirmação dizendo que os objetivos da formulação das estratégias de busca deve ser:

- recuperar um número suficiente de registros relevantes;

- evitar que sejam recuperados registros irrelevantes;
- evitar recuperar um número excessivo de registros
- Evitar recuperar um número insignificante de registros

Segundo Bastos (1994), cada diferente tipo de representação (termos, frases, citações, resumos, língua natural, texto completo, etc.) leva a distintas técnicas de recuperação e como conseqüência a diversos tipos de resultados. Uma dessas técnicas é a de coincidência exata, em que a representação do documento deve corresponder exatamente a mesma representação da pergunta. È a técnica mais adotada pelos sistemas de informações disponíveis que usam operadores lógico-matemáticos.

Outra técnica que pode ser usada é a de coincidência parcial entre os documentos e a coleção ou a coleção e o documento. Os documentos são recuperados por um grupo de características frases ou conceitos. Essa técnica usa estratégias tentam melhorar os resultados das buscas como ponderação, análise de freqüência, análise estrutural lógica e lingüística, redes semânticas etc.

Os mecanismos de busca possuem características próprias para a recuperação da informação variando de mecanismo para mecanismo. Entretanto, a maioria utiliza dois níveis de especificação de expressão de busca: básico e avançado.

O nível básico geralmente utiliza janelas e menus que fazem a busca por lógica buscas booleana ou utiliza ainda a delimitação de frases utilizando aspas.

A lógica booleana está baseada na Álgebra binária de *Boole* e na teoria dos conjuntos, possui segundo Robredo (2005), uma base binária sólida e simples e tem ampla utilização na recuperação de informações textuais.

Segundo Rowley (2002) a lógica de buscas é utilizada para ligar os termos que descrevem os conceitos presentes no enunciado das buscas, permitindo a inclusão de todos os termos relacionados e as combinações aceitáveis e inaceitáveis de termos de busca. Os operadores lógicos booleanos são: E, OU, NÃO, além de suas variações como: E NÃO.

O nível mais avançado, além de oferecer expressões booleanas mais complexas fornece também recursos mais sofisticados. Podem usar operadores de extensão que faz a busca em radicais de palavras, empregando um caractere indicativo de truncamento, um asterisco * por exemplo. Este caractere instrui o sistema a fazer uma busca numa seqüência de letras, independente dessa seqüência formar ou não uma palavra completa.

O trucamento mais importante é à direita, no qual são ignorados os caracteres situados à direita da seqüência de caracteres. O truncamento à esquerda será útil nas situações onde ocorrem diversos prefixos. (Rowley, 2004)

As buscas mais avançadas por proximidade também podem oferecer um recurso chamado de operadores meta-sintáticos. Esses operadores se baseiam na situação e na ordem que aparece os termos num documento.Podem determinar que duas palavras se encontrem uma em seguida à outra, utilizando o operador (ADJ), num mesmo campo ou parágrafo ou ainda que duas palavras estejam numa distância especificada uma da outra.

Cada mecanismo de busca aceita um grupo de operadores lógicos específicos, veja na tabela abaixo:

MECANISMOS	BOOLEANO	CONECTOR	PROXIMIDADE	TRUNCAMENTO	CAMPOS INDEXADOS	LIMITAÇÃO	DISTRIBUIÇÃO DOS SITES
Google	-, OR	and	Frase	Não. Só na busca avançada	Título, url, link, site, e outros	Language m, tipo de arquivo, data e domínio	Relevância
Yahoo!	AND, OR, NOT, ( ), -	and	Frase	Não. Só na busca avançada	Título, url, link, site, e outros	Language m, tipo de arquivo, data e domínio	Relevância
Ask	-, OR	and	Frase	Não	Título, url, link, site, e outros	Language m, n° de site, data	Relevância
Live Search	AND, OR, NOT, ( ), -	and	Frase	Não	Título, url, link, site, local	Language m, n° de site	Relevância
Gigablast	AND, OR, AND NOT, ( ), +, -	and	Frase	Não	Titulo, site, ip e outros	Domínio, tipo de arquivo	Relevância
Exalead	AND, OR, NOT, ( ),-	and	Frase, NEAR (conector)	Sim	Título, url, link, site, e outros	Language m, tipo de arquivo, data e domínio	Relevância, data
WiseNut	- only	and	Frase	Não	Não	Language m	Relevância

Tabela 2: Operadores lógicos dos mecanismos de Busca
Fonte: Search Engine Showdown

10 Limitações dos Sistemas de Busca da Web

Como descrito anteriormente cada sistema de busca possui uma forma própria de recuperação da informação. Porém a maioria são extremamente limitados principalmente ao usuário leigo. Yamaoka (2002) destaca como principais limitações às formas de recuperação:

- recuperação somente por coincidências de palavras;
- não oferecem recursos de interpretação sintática quando em frases que usam a linguagem natural
- não fazem pesquisas fonéticas;
- os recursos de busca multilingual são limitadíssimos
- não possuem tratamento semântico dos termos inseridos em uma busca;
- requer que o usuário conheça o assunto que realiza busca para a correta seleção de palavras-chaves ou frase.
- Não conseguem indexar todo o conteúdo da *web*.

Além dessas outras limitações serão descritas nos tópicos abaixo:

10.1- Restrições da busca booleana

Apesar de sua simples utilização a busca booleana apresenta algumas limitações e esse fator influencia diretamente na eficiência da recuperação da informação pelos sistemas de busca da *web*. Em estudos precursores como o de Cooper (1984 apud BASTOS, 1994, p. 27) verificou-se que o usuário tem dificuldades em manejá-los. A conjunção "e" pode confundir-se com "o" pois tem significados diferente na linguagem natural, porém funciona de maneira distinta da na linguagem do sistema.

Robredo (2005), enumera alguns problemas relacionados com os operadores booleanos, como:

- o uso dos operadores booleanos, não conseguem fazer distinção entre palavras variantes de forma, desinências e flexões afetando a qualidade da recuperação

dos documentos. Pode-se solucionar o problema utilizando a truncagem de termos.

- não há distinção da posição que o termo se encontra, o que leva à recuperação de muitos documentos irrelevantes. Esse problema pode ser amenizado com o uso do recurso de adjacência ou proximidade, já citado anteriormente.
- Outra questão é que o método booleano pode ser afetado pela existência de termos polissêmicos, ambíguos ou imprecisos. A delimitação da área de conhecimento por meio do uso de metadados e qualificadores, na descrição de registros pode amenizar o problema.

Bookstein (1985) aponta outras restrições ao uso da lógica booleana:

- perda de textos cuja representação pode corresponder parcialmente à representação da pergunta;
- não classifica os textos recuperados;
- não considera a importância de um determinado conceito no documento ou na pergunta;
- depende da coincidência perfeita entre as representações do documentos e da pergunta, corresponder exatamente ou vocabulário usado.

Bastos (1994) sugere que os métodos que incluem o processamento da linguagem natural à recuperação da informação ofereçam uma forma mais proveitosa e precisa de tratar a informação. A autora defende que o processamento da linguagem natural (PLN) é parte essencial no processo de recuperação da informação, para possibilitar a interação homem máquina na linguagem natural e para reconhecer unidades de informação que representem com maior precisão o conteúdo dos documentos através da análise de sua estrutura lingüística.

10.2- Web Oculta

A oculta é a parte da *web* que os mecanismos de busca tem dificuldade de recuperar e indexar e esta é outra limitação relacionada aos sistemas de busca. Atualmente, há muito mais na *web* do que apenas texto. Fotos, programas de computador, filmes e bancos de dados formam uma riqueza de informação que nem todos os mecanismos de buscaes estão preparados para localizar e indexar.

Sendo assim, podemos dizer que parte do conteúdo existente na *web* está mesmo "invisível", mas apenas para os mecanismos de buscas que são incapazes de encontrá-lo. Estes sites não aparecem nos resultados apresentados por estas ferramentas de busca. Estima-se que esta parte oculta da *web* tenha mais que o dobro do tamanho da parte visível e seu conteúdo é bastante relevante.

Segundo Yamaoka (2002), fazem parte da *web* oculta:

- Conteúdo de banco de dados que formam páginas dinâmicas montadas pelos usuários, como o Orkut, por exemplo;
- Conteúdos protegidos por *firewell* em redes privadas;
- Conteúdos protegidos por sites protegidos por senhas de acesso;
- Documentos isolados da web (que não recebem hiperligações de outros documentos);
- Páginas com frames e image-maps também não são indexados por alguns mecanismos de busca.

Numa visão mais abrangente, Braski (2004) diz que há, basicamente duas razões para estes *sites* estarem fora dos bancos de dados de grande parte dos buscadores:

- questões técnicas que impedem o acesso dos *spiders* a alguns tipos de sites.

- por decisão dos administradores dos mecanismos de busca.

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

Os mecanismos de busca genéricos não são capazes de acessar os conteúdos das páginas transitórias geradas por outros bancos de dados. Quando um *spider* se depara com um banco de dados é como se encontrasse uma biblioteca com portas de segurança invioláveis.São capazes de ler o endereço da biblioteca, mas não podem dizer nada sobre os livros, revistas ou outros documentos armazenados. Os robôs não têm dificuldade em encontrar a interface de um banco de dados porque se assemelham a outras páginas *web* que utilizam formas interativas. Mas, os comandos que permitem o acesso ao conteúdo do banco de dados são incompreensíveis.Os robôs não estão programados para entender a estrutura de um banco de dados, ou as linguagens utilizadas para recuperar a informação.

Por política de exclusão os mecanismos de busca limitam o número de páginas que coletam utilizando alguns critérios de tal forma que certos tipos de linguagem de programação como: Flash,*Schokwave,Word,WordPerfect* ,arquivos executáveis e comprimidos, páginas formatadas em *Portable Document Format* (PDF) etc.,podem ser excluídas porque,além de aumentarem o custo de operação das ferramentas de busca,tem menor procura.

Segundo Yamaoka (2002), fazem parte da *web* oculta:

- Conteúdo de banco de dados que formam páginas dinâmicas montadas pelos usuários, como o *Orkut*, por exemplo;
- Conteúdos protegidos por *firewell* em redes privadas ;
- Conteúdos protegidos por *sites* protegidos por senhas de acesso;
- Documentos isolados da *web* (que não recebem hiperligações de outros documentos);
- Páginas com *frames* e *image-maps* também não são indexados por alguns mecanismos de busca.

De acordo com Araújo (2001), o conhecimento da *web* oculta se justifica por dois motivos:

- é em boa parte gratuita, ou seja, está lá para usarmos quando necessário e sem custo;
- geralmente costuma ter mais qualidade e ser mais relevante em relação ao que está disponível na *Web Visível ou Superficial*.

> "Boa parte da informação da Web Oculta ou Profunda está em bancos de dados de organizações governamentais, instituições de ensino e pesquisa e constitui fonte utilíssima e de qualidade para pesquisa bibliográfica. Além disso, essa informação geralmente existe em bancos de dados específicos para determinadas áreas do conhecimento (Medicina, Psicologia, Filosofia, Engenharia etc), o que a torna mais relevante para pesquisadores dessas áreas."(ARAÙJO, 2001)

11- Aprimoramento da Recuperação da Informação na Web

11.1- Evolução dos mecanismos de busca

Diversos mecanismos de busca vem aprimorando suas técnicas, como é o caso do *Scirus* (http://www.scirus.com) um buscador especializado em pesquisas da área científica, que faz um controle rigoroso controle terminológico. Quando se faz uma busca o sistema além de trazer o resultado da recuperação, oferece uma lista de termos relacionados com a expressão de busca.

Segundo Feitosa (2006), um outro mecanismo que vem aprimorando suas técnicas de busca é o Google que em meados de 2003 introduz em seus serviços o operador semântico "~". O sistema quando faz busca, retorna também resultados que contém sinônimos e termos relacionados, infelizmente ainda não está disponível no Brasil.

11.2- Web Semântica

Conforme descrito nas seções anteriores a *web*, da forma como foi criada e utilizando a linguagem HTML se transformou em algo que qualquer pessoa pudesse manipular sem qualquer padrão ou norma, cresceu de maneira desgovernada e caótica e se apresenta atualmente como um grande repositório de documentos heterogêneos em forma e conteúdo e descritos de forma pobre, sem que se saiba ao certo o que está por traz de cada link ou página.

Diante deste cenário Berners-Lee criou o consórcio W3C já citado anteriormente, com o objetivo de corrigir as falhas de *web* inicial. Começa a surgir então o conceito *Web* Semântica como tentativa de fornecer padrões que possibilite a representação semântica do conhecimento.

A *Web* Semântica é definida por Berners-Lee como: "uma extensão da Web já existente, onde a informação se encontra de bem definida e entendível, melhorando a cooperação e a comunicação entre o homem e o computador" (BERNERS-LEE, 2006).

Um dos grandes problemas para o desenvolvimento da interoperabilidade semântica, os sistemas "conversando" entre si semanticamente, é fazer com que a realidade do processo de significação corresponda necessariamente ao signo apresentado. Para tratar esse problema é preciso considerar os metadados como uma forma de possibilitar a associação dos documentos com seus significados e as ontologias como forma de compartilhar significados em comum. Ambos buscam uma linguagem única capaz de representar conhecimentos e regras, além de inferir novos dados.

Segundo Grimaldo (2004) a *web* semântica precisa de algumas ferramentas que permitam sua construção e lhe dê estruturas necessárias:

1. Uma linguagem que estruture os objetos digitais sintaticamente, denominado XML (*eXtensible Markup Language*)
2. Um formato que estruture o significado da informação que os objetos digitais possuem (em conjunto com os metadados associados à ele) denominado RDF (*Resource Description Framework*)
3. Um programa de computador que recupere a informação existente baseado na inteligência artificial, denominado Agentes Inteligentes.

4. Um conjunto de regras que permitam aos Agentes Inteligentes mover-se dentro da web com liberdade e de acordo com o perfil informacional do usuário que o use, denominado Ontologias.

A linguagem XML assim como a linguagem HTML se originou do SGML e contêm *tags* para descrever o conteúdo do documento. Sua principal vantagem em relação a linguagem HTML é seu foco na descrição dos dados do documento, funciona como uma espécie de metalinguagem. Duas outras vantagens é sua flexibilidade de criação, expansão e uso e a possibilidade de criar etiquetas de caráter semântico de acordo com as necessidades do criador.

O padrão RDF é uma recomendação da W3C que deve vir a ser implementada na confecção de páginas da *Web* Semântica. O RDF estabelece um padrão de metadados para ser embutido na codificação XML, e sua implementação é exemplificada pelo RDF *Schema*, RDFS, que faz parte da especificação do padrão.(SOUZA, 2004). O padrão deve permitir o agrupamento dos dados com uma sintaxe e semântica única. O RDF se baseia num esquema de triplas: um sujeito, um objeto e uma ação e deve permitir que a máquina entenda a estrutura e a organização dos metadados. Dessa forma o resultado das busca nos sistemas de busca se tornarão mais preciso.

Os agentes são programas de computador capaz de coletar, processar e compartilhar com outros programas as informações da *web*. Acrescentado o termo inteligente, esse programa utilizando as técnicas de inteligência artificial, deverá ser capaz de se adaptar às necessidades de informação do homem e inferir resultados para conseguir uma resposta mais efetiva e eficaz da tanto em tempo de resposta quanto em conteúdo.

As ontologias funcionam como uma espécie de caminho que terá que ser percorrido pelos agentes inteligentes dentro da *web*. Segundo Feitosa (2006) as ontologias do ponto de vista da representação do conhecimento, não podem ser compreendidas como um vocabulário informal, ou mesmo uma linguagem de termos estruturados- como os tesauros, por exemplo-, mas que requer uma possibilidade de interpretação algorítmica dos seus significados e, por conseguinte, uma representação em uma linguagem formal, cujo processamento de significados pode ser realizados por máquinas. A utilização das ontologias também serão capazes de extrair e agregar informações de diferentes *sites* para interpretar e resolver situações.

A *web* atual ainda está longe se possuir todas as características descritas e transforma-se em *Web* Semântica, mas é importante se os caminhos comecem a ser trilhados para que talvez possamos ter um dia uma *web* parecida com o sonho de Berners-Lee.

11.3- Web Intelligence

A *Web Intelligence* (WI), de acordo com Yamaoka (2002) é um campo de estudo recente, concebido em 1999 por um grupo de pesquisadores do Canadá (*Iniversity of Regina*), Japão (*Maebashi Institute of Tecnology e Waseda University*) e Hong Kong (*Hong Kong Bapitist University*). Foi reconhecida como uma nova direção para pesquisas e desenvolvimento científico, explora a inteligência artificial e a informática avançada na *web* e *internet*. Dentre essas pesquisas estão:

- Referentes à inteligência artificial (AI)- representação, planejamento e descoberta de conhecimento na rede e mineração de dados (*data mining*), agentes inteligentes e a *Social Network Intelligence* (rede social inteligente).
- Referentes á informação tecnológica (IT)- redes de rádio, redes sociais e outros.

A *Web Intelligence* é uma das mais importantes promessas de pesquisa em informação tecnológica (IT) e sobre agentes inteligentes, recurso de importância fundamental para a *Web* Semântica.

De acordo com Yamaoka (2002), assim como na inteligência artificial (AI) os fundamento da WI pode ser estabelecida por esboços resultantes de várias disciplinas relacionadas, como por exemplo:

- Matemática: computação, lógica e probabilidade;
- Matemática aplicada e estatística: algoritmo, lógica não- clássica, teoria da decisão, teoria da informação, teoria de medições, teoria da incerteza;
- Psicologia: psicologia cognitiva, ciência cognitiva, interação homem-

máquina, interface usuário;

- Lingüística: lingüística computacional, processamento da linguagem natural, tradução automática;
- Tecnologia da informação: Ciência da Informação, banco de dados, sistemas de recuperação de informação, *Data Mining*, Sistemas especialistas, sistemas baseados em conhecimento, sistemas de suporte a decisão, agentes inteligentes de informação.

A figura abaixo representa os campos de atuação da WI:

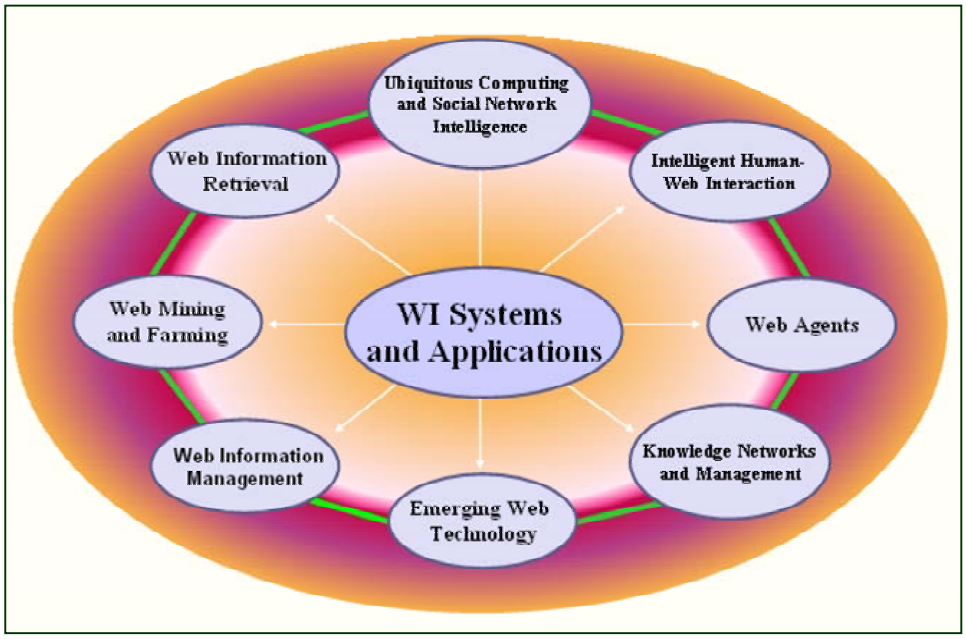

Figura 18: Campos de atuação da *Web Intelligence.*
Fonte: *Web Intelligence Consortium*

O *Web Intelligence Consortium* (WIC) (http://wi-consortium.org/) é uma organização internacional sem fins lucrativos, que se dedica a avançar na pesquisa científica e desenvolvimento industrial para a era da *web* e agentes inteligentes.

O WIC promove colaborações entre centros mundiais de pesquisa de WI e os membros dessas organizações, por meio de conferências, *workshop*, publicações oficiais de livros e jornais, boletins, relatórios oficiais e padrões para a *web*.

Dentre as maiores atividades desenvolvidas pelo WIC estão:

- Organização da *web* internacional/ regional e os agentes inteligentes por meio da realização de conferências e *workshop*;

- Publicação de jornais, livros e boletins sobre a *web;*
- Criação de centros mundiais de suporte à pesquisa científica do WIC e tecnologias de inteligência artificial.

Atualmente o WIC dispõe de quinze centros de pesquisa em diversos países como Austrália, Áustria, Pequim, França, Canadá, Japão, México e uma sede nos Estados Unidos.

12 Conclusão

Este estudo constituiu-se como uma tentativa de avançar na compreensão do funcionamento dos sistemas de busca da *web*. Conforme apresentado, o entendimento destes sistemas é bastante complexo, pois além de sua grande quantidade e variedade, estão constantemente modificando-se de forma que a literatura nem sempre consegue acompanhar.

É fundamental que bibliotecários como estudiosos da recuperação da informação comecem a aplicar seus conhecimentos no campo da informação digital, já que as iniciativas brasileiras ainda são muito escassas de estudos sobre a recuperação da informação na *web* e em projetos ligados à esse campo, como as bibliotecas digitais. Da mesma forma, a eficiência do tratamento da informação da *web* não depende somente de tecnologias, mas do uso das mesmas por parte de profissionais capacitados.

Além disso, por meio de estudos como este sobre os sistemas de busca e de estudos sobre sua interação com o usuário, podem-se criar novas interfaces utilizando-se dos avanços gerados pela *web* semântica que, por meio do formato RDF, permite a construção de sistemas de busca mais intuitivos e coerentes com o funcionamento cognitivo dos seres humanos.

Dessa forma, este estudo ao apresentar a visão geral das principais categorias de sistema de busca que a *web* dispõe na atualidade para recuperar informação e suas perspectivas futuras, serve de apoio ao profissional da informação e aos usuários da *web* em geral. O conhecimento sobre os recursos de busca proporcionam o uso mais eficaz dos sistemas e redes de informação na recuperação da informação.

Por fim, face à importância do conhecimento sobre os sistemas de busca, sugiro como investigação de pesquisa futura o aprofundamento dos estudos das estratégias de busca já que foi a partir de estudos sobre este tema que difundiu-se do emprego de buscas de melhor coincidência e vínculos de hipermídia; pesquisas orientadas para a avaliação da qualidade dos sistemas de busca da *web* e pesquisas sobre sua usabilidade.

13 Referência Bibliográfica

ALONSO ARÉVALO, Julio. **Recuperación de la información : la búsqueda bibliográfica**, 2004. Disponível em: <http://eprints.rclis.org/archive/00002521/ >. Acesso em: 27 nov.2006.

AMAT, N. **Documentación científica y nuevas tecnologías de la información**. 3. ed. Madrid: Pirâmide, 1989. 527 p.

ARAÚJO, José Paulo. **Invisível, Oculta ou Profunda?** a *web* que poucas ferramentas enxergam.[S.l]: Comunicar, 2001Disponível em: <www.comunicar.pro.br/artigos/weboculta.htm> . Acesso em: 6 nov. 2006.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 6023*: informação e documentação: referências: elaboração. Rio de Janeiro, 2002.

_____ . **NBR 6024**: numeração progressiva das seções de um documento. Rio de Janeiro, 1986.

_____ . **NBR 6027**: sumário. Rio de Janeiro, 1989.

_____ . **NBR 10520**: apresentação de citação em documentos. Rio de Janeiro, 2002.

_____ . **NBR 14724**: trabalhos acadêmicos: apresentação. Rio de Janeiro, 2001.

ASK. **About Ask**. Disponível em: <http://about.ask.com/en/docs/about/index.shtml>. Acesso em: 24 nov. 2006.

BEAL, Adriana. **Gestão estratégica da informação**: como transformar a informação e a tecnologia de informação em fatores de crescimento e de alto desempenho nas organizações. São Paulo: Atlas, 2004. p. 137. ISBN 85-224-3764-5

BELKIN, N.J. In: JONES, K.S. **Information retrieval experient**. London: Butterworks, 1981. Ineffable concepts ininformation retrieval, p. 44-58.

BELKIN, N. J. e CROFT, W. B. Retrieval tecniques. **Annual Review of Information Science and Tecnology**, [S.l], v. 22, p. 112-119, 1987.

BERNERS-LEE Tim. **The semantic web lifts off**, W3C, 2002. Disponível em: <http://www.w3.org/2001/sw/> . Acesso em: 24 nov. 2006.

BOOKSTEIN, A. Probability and fuzzy-set applications to informatio retrieval. **Annual Review of Information Science and Tecnology.** v. 20, p. 117-151, 1985.

BRAGA, Gilda Maria. Informação, ciência da informação: breves reflexões em três tempos. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 84-88, jan./abr. 1995.

BRANSKI, RM. Recuperação de Informações na *web*. **Perspectivas em Ciência da Informação**, v. 9. n. 1, p. 70-87, jan./jun. 2004.

BRITANNICA. **About us**. Disponível em: <http://corporate.britannica.com/about/ >. Acesso em: 24 nov. 2006.

_____. **Britannica subjects**. Disponível em : <http://www.britannica.com/eb/subject>. Acesso em:27 nov. 2006.

CÉNDON, Beatriz Valadares. Ferramentas de busca na *web*. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

CLUSTY. **About Clusty**. Disponível em: <http://clusty.com/about >. Acesso em: 24 nov. 2006.

COSTA, Antônio Felipe Corrêa da. **Ciência da informação:** o passado e a atualidade. **Ciência da Informação**, Brasília, v. 19, n. 2, p. 137-143, jul./dez. 1990.

DOGPILE. **About Dogpile**. Disponível em<http://www.dogpile.com/info.dogpl/search /help/about.htm>. Acesso em: 25 nov. 2006

DMOZ. **O que é o Open Directory Projet**. Disponível em: <http://www.dmoz.org/ about.html >.Acesso em: 22 nov. 2006.

FEITOSA, Ailton. **Organização da informação na *web*:** das tags à *web* semântica. Brasília: Thesaurus, 2006. 131 p. 85-7062-499-9

GOOGLE. **Tudo sobre o google**. Disponível em:< http://www.google.com.br/intl/pt-BR/about.html> . Acesso em: 22 nov. 2006.

IXQUICK. **About Ixquick**. Disponível em: <http://us.ixquick.com/eng/aboutixquick/ >. Acesso em: 24 nov. 2006.

KENT, A. **Manual da recuperação mecânica da informação**. São Paulo: Polígono, 1972.

LANCASTER, F. W. **Information retrieval systems: charactistics, testing and evaluation**. 2. ed. New York, NY: Wiley, 1978.

LANCASTER, F. W. **Indexação e resumos:** teoria e prática. 2. ed. Brasília: Brinquet de Lemos, 2004. 452 p. ISBN 85-85637-24-2

LIVE SEARCH. **Ajuda do Live Search**. Disponível em: <http://help.live.com>. Acesso em: 24 nov. 2006.

LOOKSMART. **About Looksmart**. Disponível em: <http://search.looksmart.com/>. Acesso

em: 25 nov. 2006.

LOPES, Ilza Leite de Azevedo Santos. **Proposta de critérios de qualidade para avaliação da informação em saúde recuperada nos sites brasileiros da world wide *web*.** 2006. 159 f. Tese (Doutorado em Ciência da Informação), Universidade de Brasília, Brasília, 2006.

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

MACEDO, Flávia Lacerda Oliveira de. **Arquitetura da informação:** aspectos epistemológicos, científicos e práticos. 2005. 190 f. Dissertação (Mestrado em Ciência da Informação), universidade de Brasília, Brasília, 2005.

MAMMA. **About Mamma**. Disponível em: <http://www.mamma.com/info/about.html>. Acesso em: 24 nov. 2006.

METACRAWLER. **About Metacrawler.** Disponível em: <http://www.metacrawler.com /info.metac/search/help/about.htm>. Acesso em: 24 nov. 2006.

MOYANO GRIMALDO, Wilmer. Sociedad de la Informacion : metadatos y futuro de la Internet en la recuperación de informacion de calidad. **Bibliotecas & Tecnologías de la Información,** 2004. Disponível em: <http://eprints.rclis.org/archive/00005274/>. Acesso em: 27 nov. 2006.

NOTESS, Greg. R.. **Search engine features chart**. [S.l.], 2006. Disponível em: <http://www.searchengineshowdown.com/features/ >. Acesso em: 27 nov. 2006.

_____. **Meta search engines**. [S.l], 2006. Disponível em: <http://www.searchengineshowdown.com/multi/ >. Acesso em: 27 nov. 2006.

ROBREDO, Jaime. **Documentação de hoje e de amanhã:** uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. rev. Ampl. Brasília, 2005. 409 p. ISBN 85- 905920-1-4

ROWLEY,Jenifer. **A biblioteca eletrônica.** Tradução de Antônio Agenor Briquet de Lemos. 2. ed. Brasília: Briquet de Lemos/Livros, 2002. 399 p. Segunda edição de Informática para bibliotecas; Título original: The eletronic library. ISBN 858563720X.

SARACEVIC, Téfko. Ciência da informação: origem,, evolução e relações. **Perspec. Ci. Inf.**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

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

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. 1-16, jan./abr. 2004. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-

19652004000100016&lng=pt&nrm=iso >. Acesso em: 27 nov. 2006.

TEIXEIRA, Cenidalva Miranda de Souza; SCHIEL, Ulrich. A internet e seu impacto nos processos de recuperação da informação. **Ciência da Informação**. Brasília, v. 26, n. 1, 1997. Disponível em: < http://www.ibict.br/cienciadainformacao/viewarticle.php?id=462&layout= abstract >.Acesso em: 27 nov. 2006.

TORQUE COMUNICAÇÕES E INTERNET**.Comunicação e internet.** Disponível: **<**http://www.torque.com.br/internet/historia.htm >. Acesso: 27 nov. 2006.

VIEIRA, Simone Bastos. **La recuperación automática de información jurídica** : metodologia de análises lógico-sintáctico para la lengua portuguesa.1994. 382 f. Tese (Doutorado em Ciência da Informação)- Universidad Complutense de Madrid, Madrid, 1994. VIVÍSSIMO. About Vivíssimo. Disponível em: <http:// >. Acesso em: 24 nov. 2006.

YAHOO DIRECTORY. **Ajuda do yahoo diretório**.Disponível em: <http://br.yahoo. com/info/diretorio.html>. Acesso em: 25 nov. 2006.

YAHOO SEARCH. **Ajuda da pesquisa Yahoo**. Disponível em: <http://br.search. yahoo.com/>. Acesso em: 24 nov. 2006.

YAMAOKA, Eloi Juniti. **Recuperação da informação na web:** cenário atual e perspectivas para o futuro. Brasília, [S.n], 2002. 19 p.