

## **Recuperação da Informação**

### **Introdução**

#### **Tipos de Sites de Pesquisa**

**Máquinas de Pesquisa**

**Diretórios de Assuntos**

#### **Técnicas de Busca e Indexação**

**Busca da Informação**

**Indexação da Informação**

**Google**

**Registro e Otimização de Páginas**

#### **Pesquisa na Internet**

**Operadores Lógicos ( Álgebra Booleana )**

**Operadores Especiais**

**Google**

#### **Sites de Pesquisa na Internet**

**Máquinas de Pesquisa**

**Diretórios de Assuntos**

**Origem dos Resultados**

## Introdução

A boa notícia a respeito da Internet e de sua parte mais visível, a Web, é que existem centenas de milhares de páginas disponíveis, esperando para apresentar informações sobre uma infinidade de tópicos. A má notícia é que a maioria destas páginas tem títulos pouco claros e estão localizadas em servidores obscuros. Assim, é imprescindível a existência de Sites de Pesquisa na Internet.

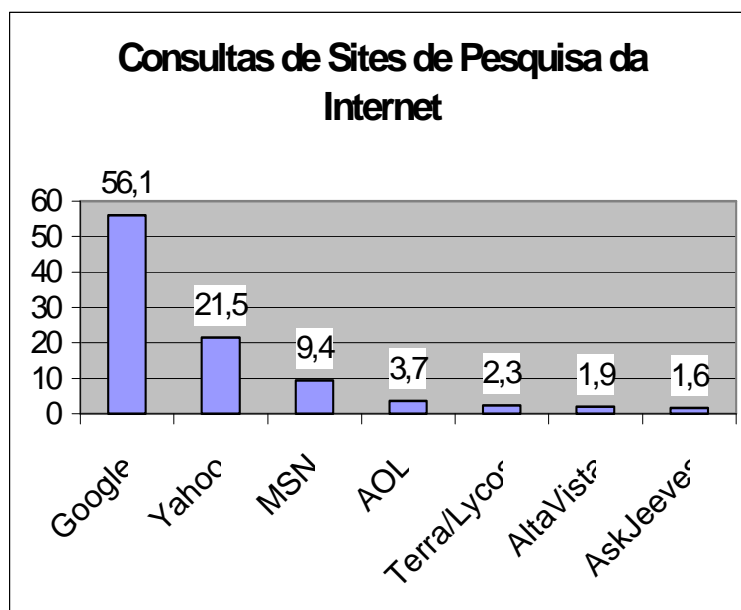
Os Sites de Pesquisa são projetados especialmente para armazenar informações sobre outros sites. Existem muitas diferenças entre eles, mas basicamente eles executam 3 tarefas:

- Eles pesquisam a Internet ou selecionam parte da Internet baseados em palavras de seu conteúdo;
- Eles mantêm um índice das palavras encontradas e onde elas foram encontradas;
- Eles permitem que usuários pesquisem palavras ou combinações de palavras nos índices criados;
- Eles permitem que usuários pesquisem por assunto.

Os primeiros sites de pesquisa mantinham índices de uma pouca centena de palavras e documentos e recebiam poucas milhares de consultas por dia. Hoje os grandes sites indexam bilhões de documentos e respondem a dezenas de milhões de consultas por dia.

Quando se fala sobre sites de pesquisa da Internet não verdade estamos falando de pesquisa na Web e eventualmente em Grupos de Discussão. Assim quando usamos o termo "Pesquisando a Internet" não estamos pesquisando a Internet, mas sim consultando um banco de dados que remete a determinadas páginas. Ao contrário, por exemplo, da classificação de livros, os documentos armazenados na Web tem grandes problemas para indexação:

- grande quantidade, ultrapassando bilhões de documentos e aumentando a cada dia;
- falta de padronização;
- grande distribuição geográfica.;
- diversidade de idiomas;
- conteúdo e formato altamente variável.



O mercado de sites de pesquisa da Internet está em grande ebulição em função do IPO (Oferta Pública de Ações) do Google no final de abril de 2004. Ao longo de 2002/2003 várias empresas foram adquiridas por outras na busca de consolidação do mercado. A Yahoo adquiriu o site Inktomi, o Overture adquiriu os sites AllTheWeb e AltaVista, e a AskJeeves adquiriu o site Teoma. Além disso a Microsoft anunciou que a próxima versão do seu sistema operacional desktop, nome de código "Long Horn", que deverá estar disponível em meados de 2006 deve conter uma ferramenta de pesquisa na Internet associada com um site da própria Microsoft. Em função disso a própria Google já anunciou um software desktop para realizar pesquisas diretamente do seu computador.

O Google é atualmente o maior site de pesquisa na Internet e os números divulgados pelo ao dar entrada em seu IPO impressionam, mais ainda por se tratar de um site onde a maioria dos usuários executa consultas gratuitamente:

- Faturamento de 2003, US\$ 961 milhões
- Lucro em 2003, US\$ 105 milhões
- Número de Funcionários: cerca de 1900
- Número de Servidores: cerca de 100
- Número de Consultas/Dia: 200 milhões
- Páginas indexadas: 4 bilhões

O site de pesquisas do Google é um dos mais completos da Internet sendo que seu principal produto é a Máquina de Pesquisa da própria Google. Além disso ele permite pesquisar Imagens, Grupos de Discussão e Diretórios de Assuntos, este último através do site "Open Directory".



## Tipos de Sites de Pesquisa e Indexação da Internet

Essencialmente existem 2 tipos de sites de Pesquisa na Internet: Máquinas de Pesquisa, Diretórios de Assuntos.

### Máquinas de Pesquisa

As Máquinas de Pesquisa (Search Engines) são sites construídos com base em dados coletados através de programas de computador escritos especificamente para esta finalidade. São baseados em técnicas de busca e indexação das palavras contidas no site. As suas características são:

- Construídos com base em programas de computador, chamados “robôs”, sem interferência humana;
- Os resultados não são organizados por categorias de assunto, mas sim através de um ranking calculado através de um algoritmo específico;
- Os sites são indexados com base nas palavras contidas em suas páginas;
- Não são avaliados no que diz respeito ao conteúdo, isso cabe ao usuário fazer;
- As pesquisas são realizadas com base em palavras, buscando palavras específicas ou combinações de palavras nas páginas do site;
- A extensão da busca e indexação alcança grande parte da Web, podendo ser enorme.

Exemplos: Google (<http://www.google.com>), AltaVista (<http://www.altavista.com>)

## **Diretórios de Assunto**

Os Diretórios de Assuntos (Subject Directories) são sites construídos com base em dados organizados e avaliados manualmente por pessoas. As suas características são:

- Construídos pela seleção humana manual, não por computadores ou programas robôs;
- São organizados em categorias hierárquicas de assuntos, divididos em páginas, porém os assuntos não são padronizados e variam muito em função do objetivo do site em questão;
- Eles NUNCA se referem ao conteúdo indexado do site, ou seja, a pesquisa pode ser feita apenas através do conceito bem geral “assunto”;
- A extensão é bem menor que nas Máquinas de Pesquisa, porém mais específica;
- Alguns sites oferecem comentários a respeito dos sites indexados.

Exemplos: Yahoo (<http://dir.yahoo.com>), Open Directory (<http://dmoz.org>)

## **Técnicas de Busca e Indexação**

As Máquinas de Pesquisa da Internet são baseadas na indexação de palavras dos sites, assim o processo de construção do banco de dados pesquisa começa com a busca das páginas e termina com a indexação das palavras.

### **Busca da Informação**

Antes da Máquina de Pesquisa poder dizer onde determinado documento está, ele precisa ser encontrado. Para encontrar informações nas centenas de milhares de páginas Web existentes as Máquinas de Pesquisa empregam um software especial chamado “Robô” ou “Spider” (Aranha), para construir uma lista de todas as palavras encontradas nos sites. O processo de construir uma lista de palavras de sites é chamado de “Web Crawling”, algo como, “Rastejamento Web”. Entretanto, para construir e manter uma lista atualizada de palavras, os Robôs de busca devem olhar uma grande quantidade de páginas.

Como os robôs pesquisam a Internet ? Usualmente, o ponto de partida é uma lista de servidores muito utilizados e páginas de Internet bem populares. O Robô começa com uma página bem popular, indexando as palavras desta página e seguindo recursivamente todos os links das páginas. Desta maneira, o sistema de busca começará a navegar pela Web, seguindo o “vento” dos links, passando pelas partes mais navegadas da Web.

O site Google iniciou como um site de pesquisa acadêmico, se transformando rapidamente no maior site de pesquisa do mundo. Quando um Robô do Google olha uma página HTML, ele considera 2 aspectos:

- As palavras na página;
- A posição da palavra na página.

Palavras que ocorrem no Título, nos Sub-Títulos ou nos META TAGS HTML e em outras posições de relativa importância são marcadas para terem uma maior consideração durante as pesquisas dos usuários. O Robô do Google foi escrito para considerar todas as palavras da página com exceção de artigos e algumas palavras curtas .

Outros Robôs podem ter diferentes abordagens. Estas abordagens usualmente tentam fazer o Robô operar mais rápido, o usuário pesquisar mais rapidamente, ou ambos. Por exemplo, alguns robôs consideram as palavras do Título, Sub-Título e Links além das 100 palavras mais usadas na página e todas as palavras usadas nas 20 primeiras linhas da página. O site Lycos usa uma abordagem parecida com esta. Outros sistemas, como o AltaVista, vão na direção contrária, indexando cada palavra da página incluindo artigos e palavras curtas

## **META TAGS HTML**

META TAGS HTML permitem guardar informações de indexação junto ao código HTML. Isto pode ser especialmente importante nas páginas onde as palavras do texto podem ter significado duplo, onde os META TAGS podem orientar o Robô sobre o significado correto das palavras. Abaixo é mostrada a estrutura típica de uma página HTML, com Cabeçalho e Corpo, juntamente com alguns META TAGS:

```
<html>
<head>
<title>Internet e Recuperação da Informação</head>
<meta name="description" content="Como funciona a Internet e como recuperar informação da Internet">
<meta name="keywords"content="Internet Informação Recuperação">
</head>
<body>
... conteúdo da página HTML ...
</body>
</html>
```

Infelizmente nem todos os Robôs de Máquinas de Pesquisa consideram os META TAGS, pois usualmente não existem e se existirem podem estar em desacordo com a página. Um META TAG entretanto é considerado, o "robots". Ele serve para informar ao Robô se este deve considerar esta página para

indexação, o que é particularmente importante em sites com geração dinâmica de páginas, que podem não existir mais quando algum usuário seguir o link de pesquisa. Abaixo é mostrado um exemplo:

```
<meta name="robots"content="noindex">
```

Existe ainda uma outra forma de orientar o Robô sobre a indexação ou não de determinadas partes do site: usando o arquivo "robots.txt", gravado no diretório raiz do site. Abaixo é mostrado um exemplo:

```
User-agent: webcrawler  
Disallow:
```

```
User-agent: lycra  
Disallow: /
```

```
User-agent: *  
Disallow: /tmp  
Disallow: /logs
```

O primeiro conjunto de linhas diz que o robô "webcrawler" pode indexar todo o site. O segundo conjunto de linhas diz que o robô "lycra" não pode indexar nada. O terceiro conjunto de linhas diz que os demais não devem indexar os diretórios "tmp" e "logs".

## **Indexação da Informação**

A tarefa de achar informação em páginas Web executada por robôs é um processo que nunca termina em função da natureza dinâmica da Web. Mas ao terminar um ciclo de busca é necessário que a Máquina de Pesquisa organize esta informação para ser pesquisada pelos usuários.

Em uma forma simples, a Máquina de Pesquisa poderiam simplesmente armazenar as palavras encontradas e as páginas nas quais foram encontradas. Uma pesquisa feita nesses dados não seria muito útil pois não indicaria a relativa importância de uma palavra, ou seja, uma palavra que fosse usada apenas 1 vez e uma que fosse usada 100 vezes, inclusive no título, teriam o mesmo peso. Assim, foi necessário desenvolver algoritmos que criem um ranking de palavras em função de parâmetros definidos no algoritmo.

Usualmente os algoritmos consideram o número de vezes que uma palavra foi usada, se foi usada no título ou em um sub-título, se foi usada no início do texto ou em um link. Cada Máquina de Pesquisa utiliza seus próprios critérios, o que faz com que pesquisas iguais gerem resultados diferentes em diferentes Máquinas de Pesquisa. O site Google, por exemplo, utiliza um critério para definir a importância de determinada página baseado em quantos outras

páginas apontam para ela. Assim, para aparecer “bem” no Google é necessário que sua página seja bem referenciada.

Assim, na etapa de Indexação, são criados os índices que permitem pesquisar rapidamente uma palavra, armazenando, além da palavra, informações que definem sua importância relativa na página.

## **Registro e Otimização de Páginas**

O **Registro de Páginas** se refere ao ato de inserir uma página em determinada Máquina de Pesquisa ou Diretório de Assuntos. A **Otimização de Páginas** se refere a ao ato de alterar as características da página de forma a melhorar sua posição na pesquisa, especialmente de Máquinas de Pesquisa.

Um site que tenha sido criado pode ter suas páginas registradas a uma Máquina de Pesquisa ou um Diretório de Assuntos. Sites como o Yahoo (<http://dir.yahoo.com>) permitem submeter gratuitamente páginas desde que não comerciais. O Google aceita URLs para indexação, mas não é garantida a data em que será indexada nem mesmo se realmente será indexada.

Os sites, é claro, oferecem serviços pagos para inclusão de páginas em suas pesquisas de forma imediata. Porém este serviço garante apenas Indexação da página e não uma boa posição na pesquisa. Para alterar a posição na pesquisa pode-se melhorar a “qualidade” da página, por exemplo, no caso do Google, solicitando a outros sites que criem links para esta página. Além disso os sites oferecem serviços pagos para inclusão das páginas em uma lista de sites pagos, que com TODA certeza aparecerão quando o usuário fizer uma pesquisa. No caso do Google, a lista dos sites pagos aparece a direita da página, em uma coluna separada.

Importante lembrar que cada Máquina de Pesquisa possui sua lógica própria para definir a importância de uma página (não considerando a lista paga, mencionada acima). Assim é importante conhecer como o site indexa para então tentar melhorar a posição da página.

## **Pesquisa na Internet**

Os sites das Máquinas de Pesquisa tem uma interface de usuário bastante comum. Usualmente é apenas uma linha de entrada de dados onde pode-se especificar quais palavras devem ser pesquisadas, além de utilizar um conjunto de operadores lógicos específicos.

### **Operadores Lógicos ( Álgebra Booleana )**

A pesquisa através do banco de dados de um site de indexação envolve a criação de uma consulta que será submetida ao site. A consulta pode ser bem simples, contendo mesmo 1 palavra, ou complexa, quando será necessário utilizar algum operador lógico Booleano:

#### **AND (E)**

As palavras ou termos unidos pelo "AND" devem aparecer em todas as páginas pesquisadas.

Alguns sites substituem o operador "AND" pelo operador "+".

A maioria dos sites considera um "AND" implícito separando as palavras, ou seja, todas as palavras especificadas devem aparecer nas páginas pesquisadas.

futebol brasileiro

*existe um AND implícito nesta pesquisa, ou seja, futebol AND brasileiro*

futebol AND brasileiro AND ronaldo

#### **OR (OU)**

Pelo menos uma das palavras especificadas devem aparecer nas páginas pesquisadas.

futebol OR brasileiro

*podem aparecer páginas com, por exemplo, "folclore brasileiro", sem a palavra futebol*

#### **NOT (NÃO)**

As palavras que seguem o operador "NOT" não devem aparecer nas páginas pesquisadas.

Alguns sites substituem o operador "NOT" pelo operador "-".

futebol NOT argentino

*não aparecerão páginas com a palavra "brasileiro", ou seja, estamos refinando uma busca simplesmente com a palavra "futebol" onde poderiam aparecer páginas de "futebol argentino"*

## Operadores Especiais

Alguns sites permitem utilizar operadores especiais que permitem considerar a posição ou proximidade das palavras:

### **FOLLOWED BY (SEGUIDO POR)**

Uma das palavras ou termos deve ser seguida da outra.

### **NEAR**

Uma das palavras deve estar a uma distância máxima de outra, por exemplo, antes das próximas 10 palavras.

*"futebol brasileiro" NEAR "arte"*  
*a palavra "arte" deve aparecer próxima a palavra "futebol brasileiro"*

### **"" (ASPAS)**

A maioria dos sites utiliza as aspas como forma de associar palavras para forma uma expressão

*"futebol brasileiro" AND ronaldo*  
*as palavras "futebol" e "brasileiro" deve aparecer como uma expressão*

## Google

O Google, além de permitir operadores Booleanos, possui uma série de operadores especiais que permitem restringir ainda mais as pesquisas. Os operadores especiais são sempre seguidos de ":", como em:

**futebol site:www.globo.com**

### **allintext:**

A pesquisa será feita apenas no texto das páginas. Deve aparecer no início da consulta.

*allintext:futebol brasileiro*

### **allintitle:**

A pesquisa será feita apenas no texto das páginas. Deve aparecer no início da consulta.

*allintitle:futebol brasileiro*

### **filetype:suffix**

A pesquisa será feita apenas em arquivos com terminação “suffix”, por exemplo, “filetype:pdf”, onde serão considerados apenas arquivos PDF.

*filetype:pdf*

**link:URL**

Serão pesquisadas as páginas que apontam para a *URL*.

*link:www.terra.com.br*

**intext: palavra**

A pesquisa será feita em páginas que contenham a palavra “*palavra*” no texto.

*futebol intext: brasileiro*

**intitle: palavra**

A pesquisa será feita em páginas que contenham a palavra “*palavra*” no título.

*futebol intitle: brasileiro*

## Sites de Pesquisa na Internet

### Máquinas de Pesquisa

Google <http://www.google.com>

O Google é o maior site de pesquisa da atualidade, com cerca de 4 bilhões de páginas indexadas, respondendo a cerca de 200 milhões de consultas por dia. Permite pesquisa via Máquina de Pesquisa, Imagens, Grupos de Discussão e Diretório, sendo o conteúdo deste último fornecido pelo site Open Directory. Além de usar os resultados para seu site, fornece pesquisas para inúmeros outros sites, inclusive o Yahoo até pouco tempo atrás.

AllTheWeb <http://www.alltheweb.com>

Máquina de Pesquisa com cerca de 2 bilhões de páginas, sendo uma segunda opção no caso do Google não resolver. Foi adquirido em 2003 pelo site Overture.

Inktomi <http://www.inktomi.com>

O Inktomi se especializou em prover conteúdo para outros sites como, por exemplo, o MSN da Microsoft. Em 2003 foi comprado pelo Yahoo para substituir o Google como fornecedor de Máquina de Pesquisa para seu site.

Teoma <http://www.teoma.com>

O Teoma utiliza uma tecnologia de indexação chamada "Subject-Specific Popularity", onde considera, da mesma forma que o Google, a popularidade de uma página em função do número de links desta página, mas, diferentemente do Google, apenas de páginas com conteúdos similares. Alega ter 1 bilhão de páginas indexadas e mais 1 bilhão parcialmente indexadas.

AltaVista <http://www.altavista.com>

O AltaVista foi uma das primeiras grandes Máquinas de Pesquisa da Internet, criado pela Digital, utilizando computadores com processadores Alpha. Foi adquirido pelo site Overture, que também adquiriu o AllTheWeb.

## Diretórios de Pesquisa

Yahoo <http://dir.yahoo.com>

O Yahoo surgiu em 1994, sendo o Diretório de Assuntos mais antigo em operação. Em 2002 ele passou a usar o Google como base de suas pesquisas via Máquina de Pesquisa. Entretanto, em função da consolidação do mercado, comprou o site Inktomi para ser independente do Google.

Open Directory <http://dmoz.org>

O site Open Directory se utiliza de editores voluntários para catalogar a web, existindo desde 1988, sendo adquirido pela AOL em 1998, fornecendo conteúdo gratuito a qualquer um. Usualmente não é utilizado diretamente para pesquisas, mas através de sites como Google, que utilizam seu conteúdo e tem uma interface bem melhor.

LookSmart <http://www.looksmart.com>

LookSmart é um site de diretórios que vende conteúdo para outros sites como o MSN Search. Além de catalogar sites comerciais mediante pagamento, utiliza o site ZEAL, que lhe pertence, para catalogar sites não comerciais de forma gratuita. Existe desde 1996.

MSN Search <http://search.msn.com>

A Microsoft vem melhorando a pesquisa em seu site MSN, sendo que atualmente utiliza os sites Inktomi, para Máquina de Pesquisa, o site LookSmart, para Diretório de Assuntos e o site Overture, para sites pagos. Entretanto, a Microsoft já anunciou que a próxima versão do Windows (Longhorn), a ser lançada em 2006, deverá conter uma ferramenta de desktop integrada a um site de pesquisa da Microsoft. Em função disso, outros sites de pesquisa, como o Google já anunciaram ferramentas desktop para integrar com seus sites.

Librarian's Index <http://lii.org>

Indexa cerca de 15000 sites cuidadosamente selecionados por profissionais especializados da área de biblioteconomia. Seu grande mérito é a qualidade dos links pelo cuidado na sua seleção e os comentários associados a cada um

## Origem dos Resultados

Os sites de pesquisa na Internet nem sempre são a fonte de todos os tipos de resultados mostrados, como mostra a tabela abaixo.

Site	Tipo	Máquina de Pesquisa	Pagos	Diretório
AllTheWeb	Máquina Pesquisa	de AllTheWeb	Overture	-
AltaVista	Máquina Pesquisa	de AltaVista	Overture	LookSmart
AOL Serch	Máquina Pesquisa	de Google	Google	Open Directory
Ask Jeeves	Máquina Pesquisa	de Teoma	Google	Open Directory
Google	Máquina Pesquisa	de Google	Google	Open Directory
HotBot	Máquina Pesquisa	de Inktomi	Overture	-
LookSmart	Diretório Assuntos	de LookSmart/Zeal	LookSmart	-
Lycos	Máquina Pesquisa	de AllTheWeb	Overture	-
MSN Search	Diretório Assuntos	de LookSmart/Zeal	Overture	-
Netscape	Máquina Pesquisa	de Google	Google	Open Directory
Overture	Pago	Overture	Overture	-
Open Directory	Diretório Assuntos	de Open Directory	-	-
Teoma	Máquina Pesquisa	de Teoma	Google	-
Yahoo	Máquina Pesquisa	de Google	Overture	Yahoo

Existem diferenças em pesquisar o banco de dados do Google através do Google e através da AOL Search, por exemplo. O assinante da AOL tem acesso a todo o conteúdo da área de assinantes da AOL indexado pelo Google. Da mesma forma, pesquisas através do Yahoo melhoram os resultados com base em informações dos diretórios de pesquisa do próprio Yahoo.