



Mecanismos de busca na web: histórico, problemas e tendências (Será a **Web Semântica** factível?)

Prof. Dr. Ailton Feitosa - CID/UnB
Grupo de Pesquisa de Bibliotecas Digitais
ailton.feitosa@gmail.com



Agenda

- Estatísticas da Internet
- Problemas na recuperação
- Histórico e desafios aos sistemas de busca
- Caracterização dos Sistemas de busca na web
- Estratégias de organização da informação na web (as gerações da web)
- Tendências
- Futuro do Profissional de Ciência da Informação



Nº mundial de *hosts*: 433.193.199

POSIÇÃO DOS PAÍSES POR NÚMERO DE HOSTS

Fonte: Network Wizards 2007

	País	Janeiro 07
1º	Estados Unidos*	255.505.524
2º	Japão (.jp)	30.841.523
3º	Itália (.it)	13.853.673
4º	Alemanha (.de)	13.093.255
5º	França (.fr)	10.335.974
6º	Holanda (.nl)	9.014.103
7º	Austrália (.au)	8.529.020
8º	Brasil (.br)	7.422.440
9º	México (.mx)	6.697.570
10º	Reino Unido (.uk)	6.650.334

Brasil

1º da América do Sul

2º das Américas

HOSTS NAS AMÉRICAS

Fonte: Network Wizards 2007

	País	Janeiro 07
1º	Estados Unidos*	255.505.524
2º	Brasil (.br)	7.422.440
3º	México (.mx)	6.697.570
4º	Canadá (.ca)	4.257.825

Usuários na América do Sul

	País	População	Usuários	%
1°	Chile	16.130.000	5.600.000	34,72
2°	Argentina	39.920.000	10.000.000	25,05
3°	Uruguay	3.430.000	680.000	19,83
4°	French Guiana	199.509	38.000	19,05
5°	Guyana	767.245	145.000	18,90
6°	Peru	28.300.000	4.570.000	16,15
7°	Brazil	188.100.000	25.900.000	13,77
8°	Venezuela	25.730.000	3.040.000	11,82
9°	Colombia	43.600.000	3.590.000	8,23
10°	Suriname	439.117	30.000	6,83
11°	Ecuador	13.550.000	624.600	4,61
12°	Bolivia	8.990.000	350.000	3,89
13°	Paraguay	6.510.000	150.000	2,30

- Usuários na Internet em 2005:
 - 1.08 bilhão
 - Projeção para 2010:
 - 1.8 bilhão
- Negócios na web em 2003:
 - 1,3 trilhão de dólares



Taxa de Permanência



Average Monthly Online Hours per Unique Visitor by Country, March 2006

Country	Avg. Hours per Visitor March 2006
Worldwide	31.3
Israel	57.5
Finland	49.3
South Korea	47.2
Netherlands	43.5
Taiwan	43.2
Sweden	41.4
Brazil	41.2
Hong Kong	41.2
Portugal	39.8
Canada	38.4
Germany	37.2
Denmark	36.8
France	36.8
Norway	35.4
Venezuela	35.3

Note: Visitors are 15 years old or older.

Source: comScore World Metrix, 2006

Top Worldwide Online Properties Among Visitors Age 15 or Older, March 2006 (000)

Country	Unique Visitors
Worldwide Total	694,260
MSN-Microsoft Sites	538,578
Google Sites	495,788
Yahoo Sites	480,228
eBay	269,690
Time Warner Network	241,525
Amazon Sites	154,640
Wikipedia Sites	131,949
Ask Network	127,377
Adobe Sites	115,774
Lycos, Inc.	109,394
CNET Networks	107,589
Apple Computer, Inc.	98,622
Real.com Network	78,104
Monster Worldwide	74,152
Wanadoo Sites	73,446

Source: comScore World Metrix, 2006



Informação na web

74,4 milhões de sites publicados

1.089.609 domínios no Brasil

1997 – 200 milhões de páginas (10 terabytes)

2000 – 1 bilhão de páginas (30 terabytes)

2004 – 10 bilhões de páginas (GUEDES,2004)

2005 – 11,5 bilhões de páginas
(GULLI;SIGNORINI,2005))



Desafios dos usuários

- Sobrecarga de informação (*information orverload*)
 - Diversos suportes
 - Diversos sistemas de busca
 - Diversos locais para buscar a informação



Desafios do usuário

- Como escrever ou especificar uma busca?
- Como interpretar a resposta fornecida pelo sistema?
- Como manipular uma resposta muito grande?
- Como priorizar os documentos?
- Como selecionar documentos relevantes?
- A solução pode ser a escolha de alguns serviços de busca que o satisfaçam





Desafios dos serviços de busca

- A web está crescendo muito mais rápido do que qualquer tecnologia atual a possa indexar;
- Grande quantidade de páginas é atualizada freqüentemente, o que força os serviços de busca a revisitá-las periodicamente;
- Dados distribuídos em diversos computadores e plataformas

Desafios dos serviços de busca

- A busca por palavras-chave pode trazer muitos resultados, nem todos relativos ao que se deseja realmente;





Desafios dos serviços de busca

- Informações não estruturadas e redundantes
- Qualidade das informações (dados falsos, desatualizados, mal escritos, mal digitados, spam)

Desafios dos serviços de busca

Web 2.0:

Editores de textos, planilhas, apresentações, chats, blogs, agendas permitirão a publicação de páginas sem adoção de padrões



Sistemas de Classificação/Indexação participativa - folksonomias

- <http://del.icio.us>
- Sistema de armazenamento de sites favoritos
- Permite que o usuário utilize tags para atribuir palavras-chaves aos documentos
- Flickr – compartilhamento de fotografias





Desafios dos serviços de busca

- Internet invisível (Web invisível ou deep web):
- Páginas que contêm frames, image-maps, animações em Flash
- Páginas dinâmicas (dados armazenados em bancos de dados)



Desafios dos serviços de busca

- O acesso à web invisível ou profunda se dá por:
 - Comunidades
 - Assinatura
 - Sistemas inteligentes



Tipologia dos sistemas de busca

- Diretórios
- Mecanismos de busca direta
- Mecanismos de metabusca



Diretórios

- Foram a primeira solução proposta para organizar e localizar os recursos da web;
- Organizam os sites que compõem suas bases de dados, quase sempre, hierarquicamente por assunto;
- Possuem equipes de especialistas em informação que selecionam e organizam os sites em hierarquicamente e em categorias;
 - Diretórios temáticos;
 - Diretórios de ferramentas de busca.
 - Classificação passível de erros
 - Documento em mais de uma área do diretório
 - Inexistência de certas áreas



Mecanismos de busca direta

- Surgimento: Archie, 1990 - Alan Emtage - Universidade McGill em Montreal.
- Servidores Gopher
- Busca por arquivos
- São repositórios eletrônicos de informações em que a ordenação depende de algoritmos



Mecanismos de busca direta

- Abrangência mais importante que seletividade
- Uso de robôs (rastejadores, aranhas);
- Critérios para determinar o nível de relevância dos sites (dependem de cada busca):
 - Popularidade dos links;
 - Localização e frequência de ocorrência dos termos;
 - Tamanho do documento (densidade);
 - Coincidência do maior número de termos no mesmo documento;
 - Links patrocinados



Estratégias de Pesquisa avançada

- A estratégia de pesquisa avançada foi introduzida pelo AltaVista e evoluiu, com o tempo, culminando no uso de formulários que permitem filtrar: tipos de arquivos, datas, domínios, idiomas, arquivos de multimídia, entre outras opções.



Mecanismos de metabusca

- Metabuscadores/Metamotores/Multibuscadores:
- Permitem a execução de uma mesma busca em mais de uma ferramenta ao mesmo tempo - de 6 a 10 normalmente - exibindo os resultados em uma só lista;
- Gasto de mais tempo para obter o resultado, dependendo do assunto;

Mecanismos de metabusca

- Possuem bases de dados menos completas;
- Indicados quando há carência de respostas em um só motor;
- Recursos de refinamento da busca ficam inacessíveis na interface do metamotor;
- Menor precisão nos resultados;





Novas formas de recuperação

- Agrupamento de documentos, conforme suas propriedades (*clustering*);
- Controle terminológico



Novas formas de recuperação

- Em busca da semântica:
 - Categorização;
 - Expansão;
 - Refinamentos

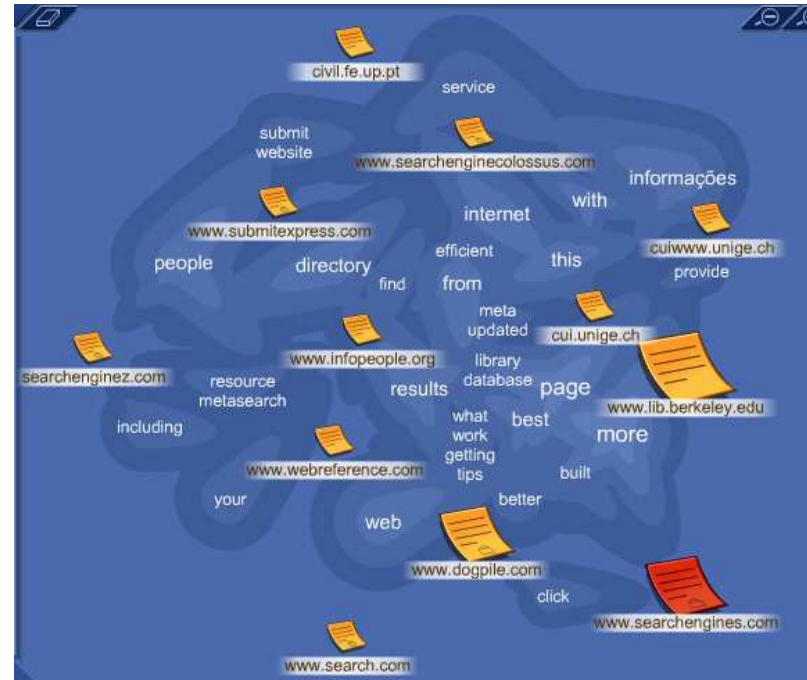


Novas formas de recuperação

- Busca por proximidade com o uso de operadores em formulários:
 - near
 - not near
 - followed by
 - not followed by
 - sentence
 - far

Nova apresentação de resultados

- Vivíssimo
 - Kartoo
 - Scirus
 - Grokker





Web 3.0: Padrões de Metadados

- Web 1.0: Uso de meta-tags (description; keywords; link);

Na web 2.0,
outras iniciativas

- Dublin Core
- TEI
- AACR2
- MARC
- GILS

Web 3.0: Web Semântica



3ª Geração – Web 3.0

- significado separado da estrutura
- baseada em RDF(S), Topic Maps, DAML+ OIL, SHOE, XOL, OWL

2ª Geração – Web 2.0

- estrutura separada da apresentação
- baseada em XML, XSL

1ª Geração – Web 1.0

- apresentação separada da localização
- baseada em HTML, PDF, CSS

- Ontologias;
- Linguagens derivadas da XML;
- Redes de relacionamentos entre termos



Web 3.0: Web Semântica

1ª Geração

```
<p><b><font face="arial">Conteúdo</font></b></p>
```

2ª Geração

```
<xml>
```

```
  <colecacao>
```

```
    <livro>
```

```
      <titulo>Dicionário de CI</titulo>
```

```
      <autor>Murilo Bastos da Cunha</autor>
```

```
      <isbn>85-85637-17-X</isbn>
```

```
    </livro>
```

```
  </colecacao>
```

```
</xml>
```

Web 3.0: Web Semântica

3ª Geração

Namespace(p =

<<http://seudominio.tipo.pais/ontologias/pessoa#>>)

Class(p: **motorista_de_ônibus** complete
intersectionOf(p: **pessoa** restriction(p: **dirige**
someValuesFrom (p: **ônibus**))))

Class(p: **motorista** complete intersectionOf(p: **pessoa**
restriction(p: **dirige** someValuesFrom (p: **veículo**))))

Class(p: **ônibus** partial p: **veículo**)

ObjectProperty(a: **é_dirigido_por** inverseOf(a: **dirige**))

Web 3.0: Web Semântica

- Um **motorista_de_ônibus** é uma **pessoa** que **dirige** um **ônibus**.
- Um **ônibus** é um **veículo**.
- Se um **motorista_de_ônibus** **dirige** um **veículo**, então tem de ser um **motorista**
- Um **veículo é_dirigido_por** um **motorista**

As inferências baseiam-se no uso de **ontologias**
– conjuntos de definições e regras de definições
relativas a um certo domínio do conhecimento
(ou de uma atividade técnica)



Web 3.0: Web Intelligence

- Data Mining
- Webwarehouse
- Feromônios;
- Web services.





Tendências: velhas soluções, nova roupagem

- AJAX Asynchronous JavaScript and XML
- Permite a criação de páginas interativas com o uso de JavaScript e XML
- Exemplo de uso em SE: Google Suggest



Tendências: indexação de mapas

- Surgimento de diversos serviços de localização por mapas
 - Google Earth
 - Google Maps
 - Yahoo Mapas



Tendências: busca por voz

- Google 2005: patente para um sistema de busca ativado pela voz.
- Uso de um modelo de linguagem, dicionário fonético e modelos acústicos, um servidor gera uma hipotética lista de palavras ou grafias faladas.



Tendências: busca por voz

- conceito possível desde que a IBM promoveu a tecnologia como um jeito de ir além das limitações dos telefones celulares, em 1999.
- Profa. Dra. Meirav Taieb-Maimon do Departamento de Sistemas de Informação da Universidade Ben-Gurion em Israel desenvolveu o sistema Maestro, que converte palavras faladas para texto;



Tendências: busca por voz

- A idéia é converter os pedidos de buscas falados em uma lista de palavras em ASCII que podem ser relacionados para um sistema de busca, com resultados listados pela ferramenta do serviço de busca.



Tendências: busca por voz

- Ao receber a frase de busca por voz de um usuário, o sistema deriva uma ou mais hipóteses de reconhecimento, cada uma associada a um peso, da frase de busca falada, e constrói uma busca booleana usando a hipótese de reconhecimento fornecendo, em seguida, os resultados ao usuário.
- Provavelmente será uma boa estratégia para a realização de buscas via celular.



Tendências: indexação de imagens

- Vídeos
 - MPEG-7 (padrão para descrição de conteúdos de multimídia com uso de metadados)
 - Baseia-se em XML
- Imagens estáticas
 - Fotografias
 - Imagens vetoriais

Desafios futuros

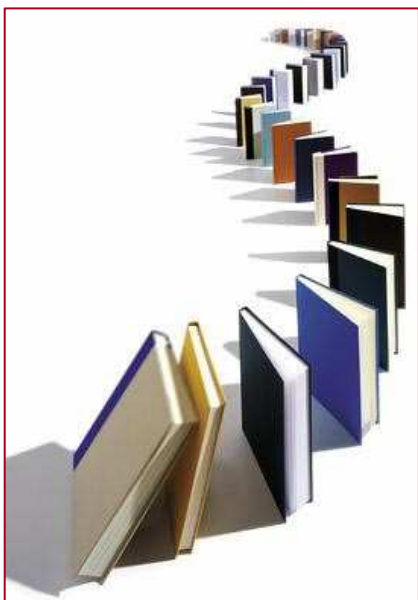
- Semântica x Multimídia
 - Textos
 - Imagens Estáticas
 - Vídeos
 - Arquivos de som
 - Acessibilidade





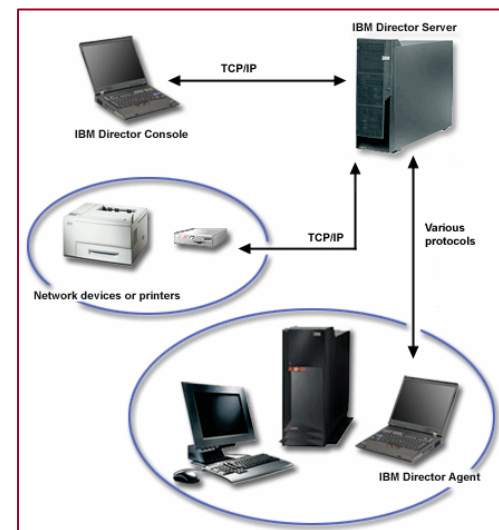
Histórias recorrentes

Sistemas de Organização



Acervo Informacional

- ✓ Classificação
- ✓ Vocabulários Controlados
- ✓ Tesouros
- ✓ Bancos de Dados Terminológicos
- ✓ Padrões de Metadados
- ✓ Ontologias (Web Semântica)



Sistemas de Recuperação

Formação e adaptação do Profissional de CI



Desorganização: mau uso de metadados

```
<meta name="generator" content="Plone -  
http://plone.org" />
```

```
<meta content="Portal para implementacao do site da  
Instituição"
```

```
name="DC.description" />
```

```
<meta content="augusto" name="DC.creator" />
```

```
<meta content="2004-02-20 11:37:06"
```

```
name="DC.date.created" />
```

```
<meta content="2004-02-20 11:37:06"
```

```
name="DC.date.modified" />
```

```
<meta content="Plone Site" name="DC.type" />
```

```
<meta content="text/html" name="DC.format" />
```



Contatos: ailton.feitosa@gmail.com

Última publicação:



- Conteúdo abordado:
 - Crescimento da Internet;
 - Elementos de Organização da Informação;
 - Busca na Web;
 - Metadados;
 - Gerações da web;
 - Web Semântica e
 - Ontologias

Referências

- CLAY, Bruce. Search Engine Chart. Disponível em [<http://www.bruceclay.com/searchenginechart.pdf>]. Acesso em 19/09/2004.
- CUNHA, Murilo B. da. Para saber mais: fontes de informação em ciência e tecnologia. Brasília. Briquet de Lemos Livros, 2001. 168 p.
- Curso de pesquisa de informação na Internet. Disponível no site Penso, logo encontro. [<http://users.skynet.be/penso.logo.encontro/curso/curso.htm>]. Acesso em 16/09/2004.
- GUEDES, Maurício F. Reinvente seu negócio através da web. Disponível em [http://www.ondeir.rec.br/artigos/mfg/artigo_040706.asp]. Acesso em 28/09/2004.
- INTEGRATED RESOURCE MANAGEMENT. Search Engine Optimization, Guidelines, Tips. Publicado em 02/09/2004. Disponível em [<http://www.integratedresourcempmt.com/searchengineneeds.html>]. Acesso em 20/09/2004.



Referências

- JOHN WILLEY & SONS INC. A history of search engines. Disponível em [<http://www.wiley.com/legacy/compbooks/sonnenreich/history.html>] . Acesso em 19/09/2004.
- LOH, Stanley; GARIN, Ramiro Saldaña. Web Intelligence – inteligência artificial para descoberta de conhecimento na web. Disponível em [<http://inf.unisinos.br/~renata/cursos/topicosiv/WebIntelligence.pdf>]. Acesso em 12/04/2006.
- MOURA, Ana M. de C. A Web Semântica: Fundamentos e Tecnologias. In: Anais do VI Congresso Internacional de Ciencias de la Computación, La Paz, pp. 46-82, out. 2001. Disponível em [<http://www.ipanema.ime.eb.br/~anamoura/publicacoes.html>]. Acesso em 16/09/2004.
- NOTESS, Greg R. Search Engine Features Chart. Disponível em [<http://www.searchengineshowdown.com/features>]. Acesso em 16/09/04.



Referências

- NUTCH. Sobre. Disponível em: [<http://www.nutch.org/docs/pt/>]. Acesso em 19/09/2004.
- SEOCONSULTANTS.COM. History of Search Engines and Directories - Search Engine History. Disponível em: <<http://www.seoconsultants.com/search-engines/history/>>. Acesso em 19/07/2004.
- SULLIVAN, Danny. comScore Media Metrix Search Engine Ratings. Publicado em 23/07/2004. Disponível em [<http://searchenginewatch.com/reports/article.php/2156431>]. Acesso em 20/09/2004. (A)
- SULLIVAN, Danny. Major Search Engines and Directories. Publicado em 28/04/2004. Disponível no site Search Engine Watch [<http://searchenginewatch.com/links/article.php/2156221>]. Acesso em 13/09/2004.(B)



Referências

- SULLIVAN, Danny. Who Powers Whom? Search Providers Chart. Publicado em 23/07/2004. Disponível no site Search Engine Watch [http://searchenginewatch.com/reports/article.php/2156401]. Acesso em 20/09/2004.(C)
- UNICAMP. Minicursos virtuais: Busca de informações na web. Disponível em [http://www.ead.unicamp.br/minicurso/bw/index.html]. Acesso em 16/09/2004.
- WALL, Aaron. History of Search Engines & Web History. Disponível em: [http://www.search-marketing.info/search-engine-history/index.htm]. Acesso em 19/09/2004.



Sites Sugeridos

- www.searchengineshowdown.com
- www.searchenginewatch.com
- www.notess.com
- <http://www.ced.ufsc.br/bibliote/virtual/busca.html>
- <http://www.search-marketing.info>

