

Inovações em Ferramentas de Busca

- *Andréa Glock* -

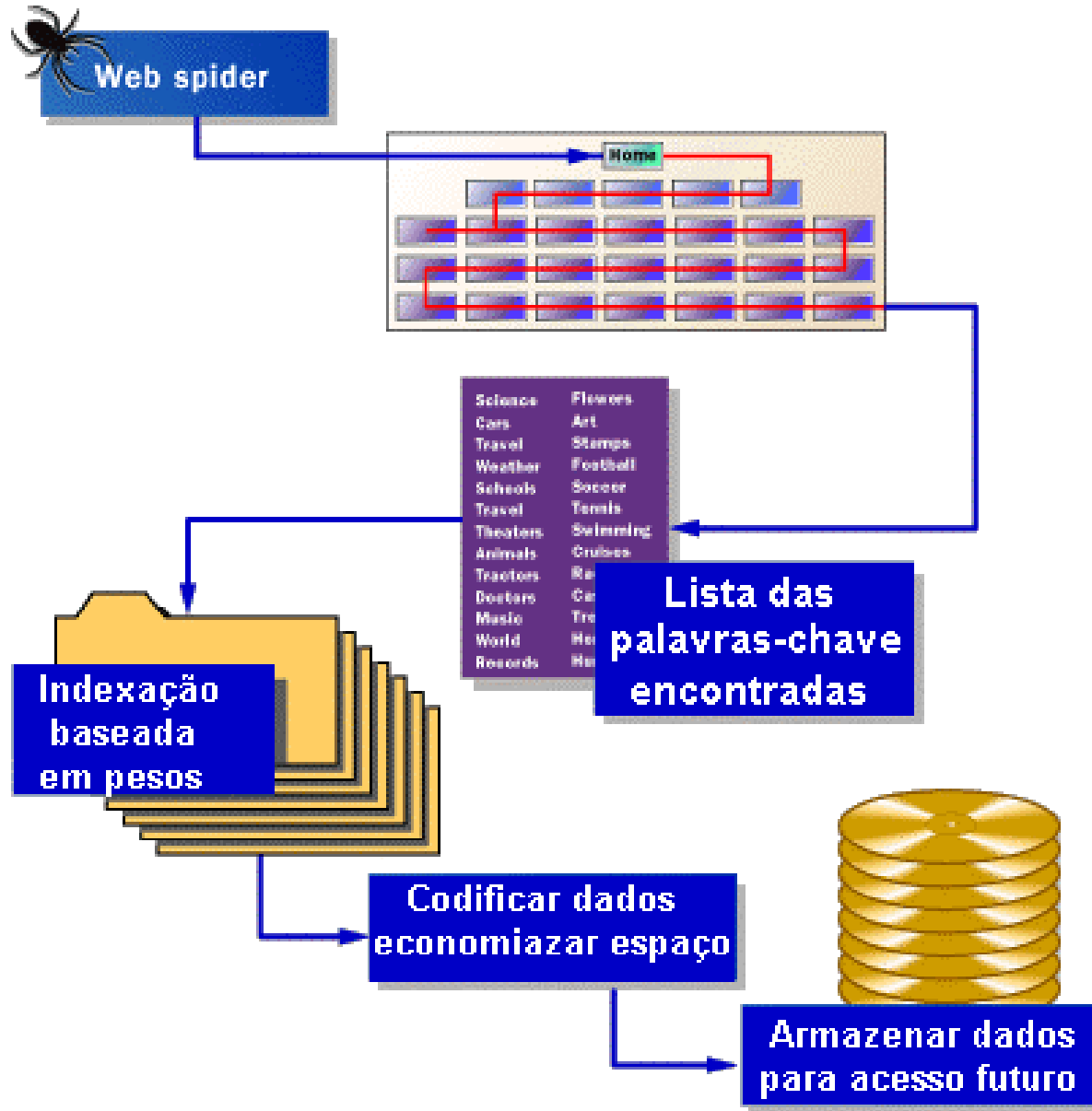


O que é um Sistema de Busca?

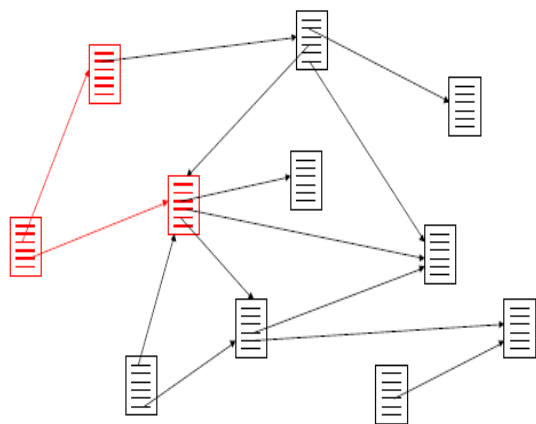
- **Conjunto de computadores, índices, bases de dados e algoritmos reunidos com o objetivo de:**
 - analisar e indexar as páginas e outros arquivos da *web*;
 - armazenar os resultados dessa análise e indexação numa base de dados;
 - recuperar e fornecer os resultados da pesquisa que preenchem os requisitos do usuário.
- **Sistema de Busca engloba as duas categorias:**
 - **diretórios** (a indexação das páginas é realizada por humanos)
 - **motores de busca** (utiliza programas de computador para a indexação as páginas)



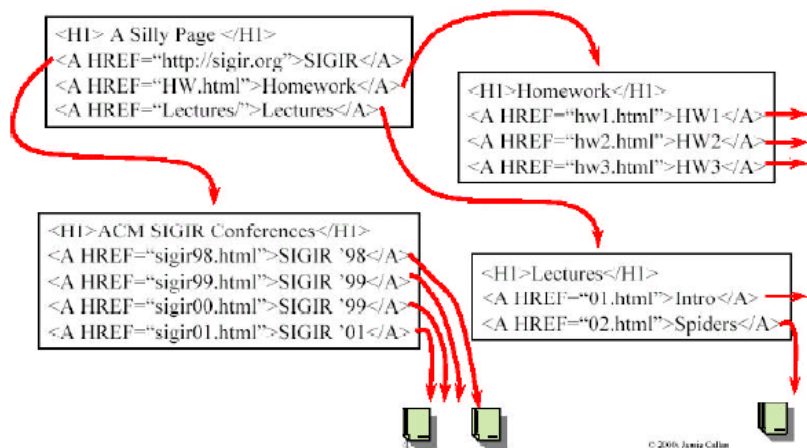
Motores ou Ferramenta de Busca



Coletores de páginas – robôs ou *sipiders*



- "visitam" as páginas (a partir de uma raiz)
- seguem os *links* contidos nesta página
- escalonador define algoritmos de seqüência da coleta
- cria uma cópia do texto contido nas páginas e guarda suas URLs



Seqüência da coleta de páginas

- Em Profundidade ou LIFO
 - Resultada em uma coleta direcionada
 - Mais páginas por *site*
 - Resultados imprevisíveis
 - Pode-se limitar o número de níveis
- Em Largura ou FIFO
 - Produz uma coleta mais abrangente
 - Visita um maior número de *sites*
 - Mais usada por ser simples de implementar



O que é coletado pelos robôs

- Documentos HTML
- Novas Ferramentas coletam formatos:
 - Imagem
 - Vídeo
 - Gráficos
 - Arquivos PDF
 - Mensagens de grupos de discussão
 - Sites de FTP



- Páginas que exigem senha de acesso
- Páginas que não tem seu link citado por outras páginas
- Páginas que contenham o metatag Meta Robot "noindex"
- Além da Web Invisível



Exemplos de programas robôs:

- Mercator (1999)
- GoogleBot (1998)
- The Internet Archive crawler (1997)



Indexação

- **Um índice é uma coleção de termos retirados dos documentos com ponteiros para os lugares onde as informações sobre os documentos podem ser encontradas.**
- **Finalidade de encontrar a informação o mais rápido possível.**



Construção de um índice

◆ Tokenização

- ◆ Os termos (palavras) da URL são separados
- ◆ São extraídos os *Tokens* = seqüências não vazias de caracteres (excluindo espaços, pontuações etc)
- ◆ O texto é transformado em um vetor através de algoritmos próprios

1 6 12 16 18 25 29 36 40 45 54 58 66 70

|That house has a garden. The garden has many flowers. The flowers are beautiful



Construção de um índice

◆ Arquivo Invertido

- ◆ Lista dos termos que aparecem no documento com os ponteiros com a posição de cada termo no texto
- ◆ Implementações típicas: B-Trees e Hashes

1	6	12	16	18	25	29	36	40	45	54	58	66	70
That house has a garden. The garden has many flowers. The flowers are beautiful													

Vocabulário

Ocorrências

beautiful	70
flowers	45, 58
garden	18, 29
house	6



Construção de um índice

◆ Stopwords

- ◆ Palavras “proibidas” nos índices (preposições, artigos etc)
- ◆ Aparecem várias vezes nos documentos mas são irrelevantes para a busca
- ◆ Normalmente não são indexadas: reduz espaço e melhora a performance da indexação
- ◆ São substituídas por marcadores de posição

◆ Stemming

- ◆ Termos considerados como derivações de um radical único
- ◆ Ocorre a remoção de inflexões, tempos verbais, gênero, número etc
- ◆ Ex: “viajar” e “viajante” são derivações de “viagem”
- ◆ Utiliza análise morfológica (ex., Porter's algorithm)
- ◆ Busca em tesouros e dicionários apresentando sinónimos e hierarquias
- ◆ Pode aumentar a abrangência da recuperação mas pode prejudicar a precisão



Algoritmos de indexação

• Vetorial

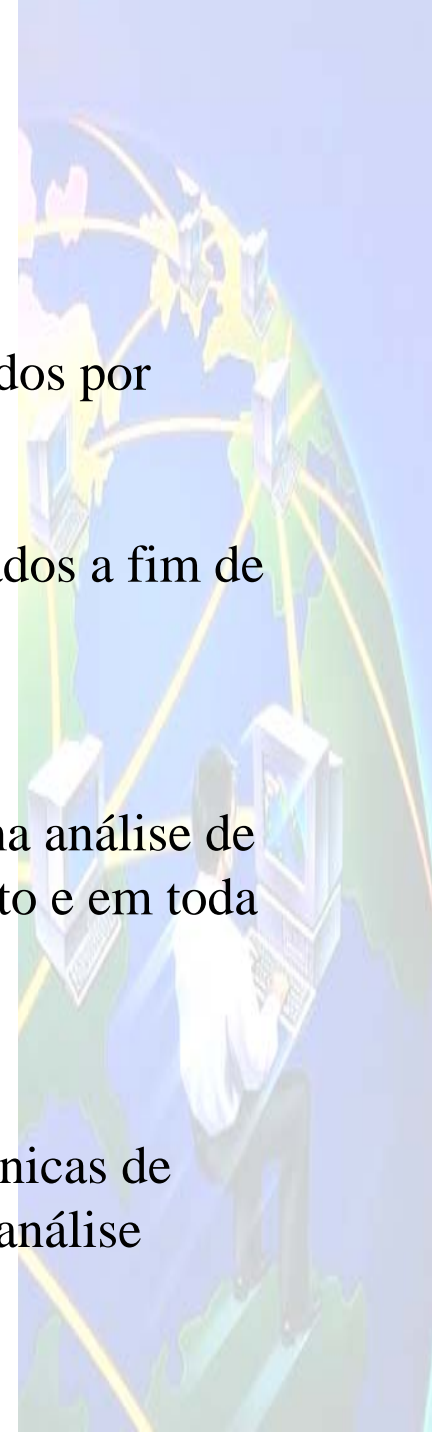
- Todos os componentes da indexação são representados por palavras
- Os componentes são vistos como vetores
- Algumas técnicas atribuem pesos aos termos indexados a fim de estabelecer sua relevância no contexto

• Estatístico

- Os termos de indexação são extraídos a partir de uma análise de frequência das palavras ou frases em cada documento e em toda a fonte de informação

• Linguístico

- Os termos de indexação são extraídos utilizando técnicas de processamento da linguagem natural, por exemplo, análise morfológica, lexical, sintática e semântica



Atribuição de pesos na indexação

- **O peso de um termo pode ser calculado através da frequência com que esse termo aparece no documento**



Maior frequência

Maior relevância



Atualização dos índices

- **Atualizam seus índices mensalmente, podendo ser feito 1x por semana**



Base de Dados

- Local onde ficam armazenadas as cópias das URLs ou endereços das páginas HTML (efetuadas pelo robô), títulos, resumos, tamanho e as palavras contidas nos documentos e os índices.
- A Base de Dados fica no servidor da ferramenta de busca
- Quanto maior o tamanho da base, melhor para a ferramenta de busca



Processamento, seleção e recuperação

- Utilizam algoritmos sofisticados para percorrer a Base de Dados da Ferramenta de Busca
- Durante o processamento devem ser calculadas as similaridades entre o vetor da consulta e os vetores de cada documento da BD
- Os pesos dos documentos são calculados na medida em que as listas invertidas são processadas
- Linguagens para consulta na Web:
 - WebSQL
 - WebOQL
 - StruQL
- Processo de *Matching* entre a expressão consultada e a informação recuperada



Métrica de avaliação

- **Objetivo:** avaliar o ranking produzido por sistemas de RI
- **2 métricas principais:**
 - Precisão (*precision*): avalia se os documentos recuperados são todos relevantes
 - Revocação (*recall*): avalia se todos os documentos relevantes foram realmente recuperados

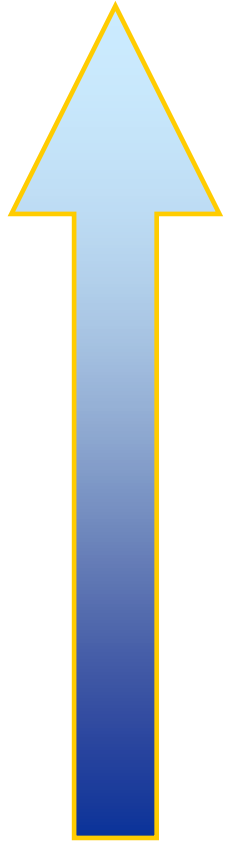


Critérios de Ordenação dos Resultados

- ***Metatags* de palavra-chave e descrição**
- **Popularidade dos *links***
- **Direct Hit**
- **Inclusão do *site* em diretórios**
- **Conceitos**
- **Pagamentos**



Evolução da Ferramentas de Busca



7ª Geração: **Ontobuscadores**..... – Ontoweb...

6ª Geração: **Multi Visão de Arquivos** – A9

5ª Geração: **Page Rank** – Google

4ª Geração : **AllTheWeb** – resultados mais organizados

3ª Geração: **Meta Buscadores** (Miner's) – MetaMiner

2ª Geração: **Robôs** (spiders) – AltaVista

1ª Geração: **Diretórios** – Yahoo





Web [Imagens](#) [Grupos](#) [Diretório](#) [Notícias](#) [mais »](#)

nanotecnologia

Pesquisar

[Pesquisa avançada](#)

[Preferências](#)

Pesquisar: a web páginas em português páginas do Brasil

Web

Resultados 1 - 10 de aproximadamente 777.000 para **nanotecnologia** (0,06 segundos)

[Com Ciência - Nanociência & Nanotecnologia](#)

Não é só na Embrapa, entretanto, que se faz **nanotecnologia** no Brasil. ... A **nanotecnologia** é extremamente importante para o Brasil, por que a indústria ...

www.comciencia.br/reportagens/nanotecnologia/nano10.htm - 30k -

[Em cache](#) - [Páginas Semelhantes](#)

[Com Ciência - Nanociência & Nanotecnologia](#)

Assim, a nanociência ea **nanotecnologia** visam, respectivamente, a compreensão eo ...

Resumindo, a **nanotecnologia** será uma revolução tecnológica de grande ...

www.comciencia.br/reportagens/nanotecnologia/nano17.htm - 34k -

[Em cache](#) - [Páginas Semelhantes](#)

[[Mais resultados de www.comciencia.br](#)]

[NANOTEKNOLOGIA](#)

A **nanotecnologia** só existe hoje como prática porque, há quase sessenta anos, ...

MINIMAPA Na hora de armazenar informações, a **nanotecnologia** pode encolher ...

www.geocities.com/capecanaveral/7754/nano.htm - 17k - [Em cache](#) - [Páginas Semelhantes](#)

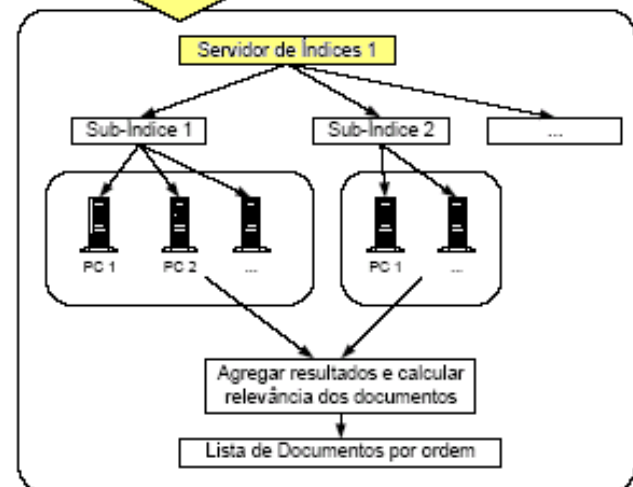
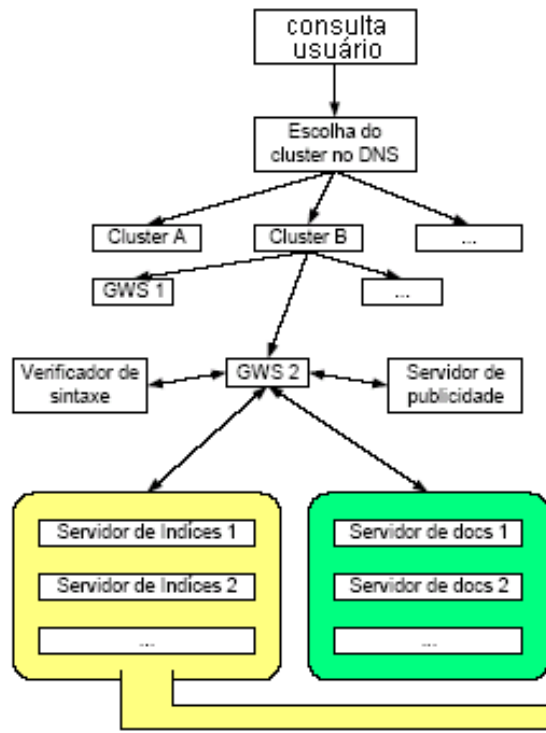
[Nanotecnologia - Inovação Tecnológica](#)

O DVD "**Nanotecnologia**: futuro", é uma iniciativa de pesquisadores da universidade ... A

situação não é diferente na **nanotecnologia**, mais especificamente, ...

www.inovacaotecnologica.com.br/noticias/assuntos.php?assunto=nanotecnologia - 16k -

Google™ Funcionamento





www.a9.com



Search: leonardo da vinci

GO

Home Prefs Toolbar Sign Out

[Advanced Web Search](#)

Hello Andra Glock. [Click here](#) if this is not you.

Hide Column Choices

- Web
- Movies
- Your Diary
- Books
- People
- Your Bookmarks
- Images
- Blog Search
- More Choices...
- Yellow Pages
- Wikipedia
- Reference
- Your History

NEW: You need to [join the A9 Instant Reward program](#) to get 1.57% off almost everything on Amazon.com.

Web Results

[\[full\]](#) [\[close\]](#)

Showing 1 - 10 of about 2,170,000

[The Da Vinci Code - Movie](#)

The Most Anticipated Thriller of our Time! Be Part of the Phenomenon
www.SoDarkTheConOfMan.com

[Leonardo da vinci](#)

Use Ask Jeeves to get the latest info on your favorite celebrities.
www.ask.com

[Leonardo da Vinci: Scientist, Inventor, Artist](#)

The Museum of Science presents an online exhibition with biography, portrait, examples of **da Vinci's** work and background information on the Renaissance.

[New] <http://www.mos.org/leonardo/> - 9k Cached [\[Site Info\]](#)

[Leonardo Home Page](#)

Leonardo da Vinci can inspire your class! Participating in activities

Book Results

[\[full\]](#) [\[close\]](#)

Showing 1 - 10 of about 15,419



[Leonardo's Notebooks](#)

by Leonardo da Vinci and H. Anna Suh (01 August, 2005) - Black Dog & Leventhal Publishers



[The Da Vinci Notebooks](#)

by Leonardo Da Vinci (04 August, 2005) - Profile Books Ltd



[Leonardo Drawings](#)

by Leonardo , da Vinci () - Editorial Benei Noaj

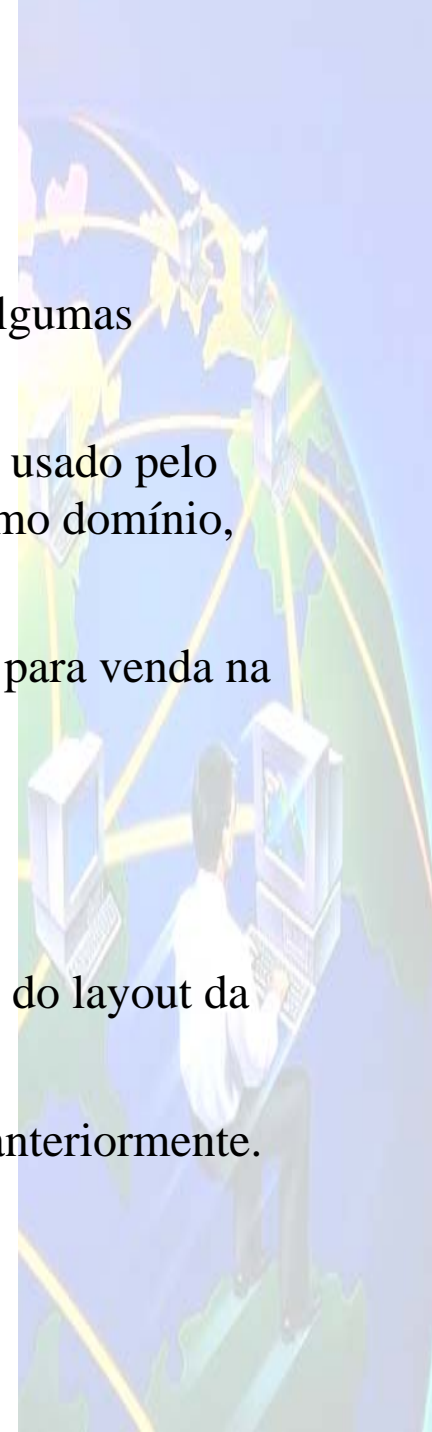
Image Results

[\[full\]](#)
[\[close\]](#)





- Da Amazon.com – criado em 2003
- Sistema de pesquisa do Google, porém substitui o *page rank* por algumas ferramentas do Alexa Web Search
- O algoritmo de ordenação dos resultados é totalmente diferente do usado pelo Google pois elimina referências que se repetem dentro de um mesmo domínio, enxugando e otimizando os resultados
- Também vasculha o interior de mais de 120 mil livros disponíveis para venda na Amazon, tudo na mesma tela
- Traz na busca textos, imagens e livros da amazon relacionados....
- Faz busca em cima de imagens
- Busca mais personalizada, cria históricos de visitas, permite ajuste do layout da tela de pesquisa trocando cores e fontes
- Através de *login*, você pode consultar todas as pesquisas já feitas anteriormente. E ainda te informa qual foi a última vez que você visitou esse *site*
- Interface é criativa, limpa e funcional





Arts
Business
Computers
Games
Health
Home

Kids and Teens
News
Recreation
Reference

Shopping
Society
Sports
World

Feedback

Enter

By Time Period>15th Century
Visual Art>Artists
Alphabetical Listing>D
Alphabetical Listing>L
Alphabetical Listing>L
Science>People
History>People
Movements>Renaissance
Alphabetical Listing>V
Alphabetical Listing>V

Da Vinci, Leonardo

Dadd, Richard
Dali, Salvador
Darger, Henry
David, Jacques Louis
Degas, Edgar
Delacroix, Eugène
Delaunay, Robert
DeMille, Jean
Demuth, Charles
Dicksee, Sir Frank
Dix, Otto
Doesburg, Theo van

Feedback

Enter your search here

leonardo da vinci

GO

or click categories above

Print Results

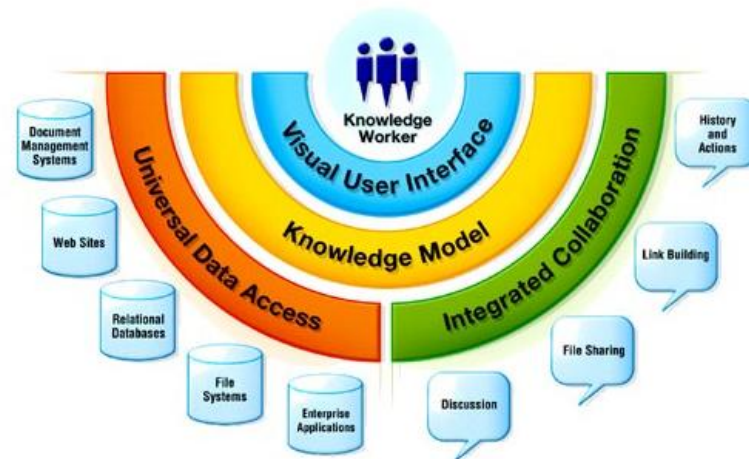
WebBrain Category Matches (1 - 3 of 3)

1. Arts>Art History>Artists>Alphabetical Listing>D
[Da Vinci, Leonardo](#)
2. World>Deutsch>Kultur>Bildende Kunst>Künstler und Künstlerinnen>Alphabetical Listing>D
[da Vinci, Leonardo](#)
3. World>Italiano>Arte>Arti Visive>Pittura>Pittori>Alphabetical Listing>D
[Da Vinci, Leonardo](#)

WebBrain Site Matches (1 - 17 of 98)

1. [LEONARDO](#)
Evaluation of the vocational training programme "Leonardo da Vinci"
<http://www.ntnu.no/intersek/leonardo/valorization/eng.htm> [Society>Government>Multilateral>Regional>European Union>Policies and Programs>Research and Education, Training and Youth](#)

- Mecanismo de busca que utiliza também uma interface renovadora baseada em diretórios e subdiretórios dispostos de forma que sejam facilmente consultados
- Ao invés de obter longas listas de categorias, o Webbrain proporciona uma pesquisa com visualização gráfica destas categorias
- O Webbrain mostra graficamente as relações de hierarquia e associações que existem entre as páginas
- Estrutura visual interessante
- Banco de dados ainda com poucos recursos.



liveplasma music movies

Search: the beatles

Artist / Band

Movie Director Actor

LAST MAPS

DISCOGRAPHY

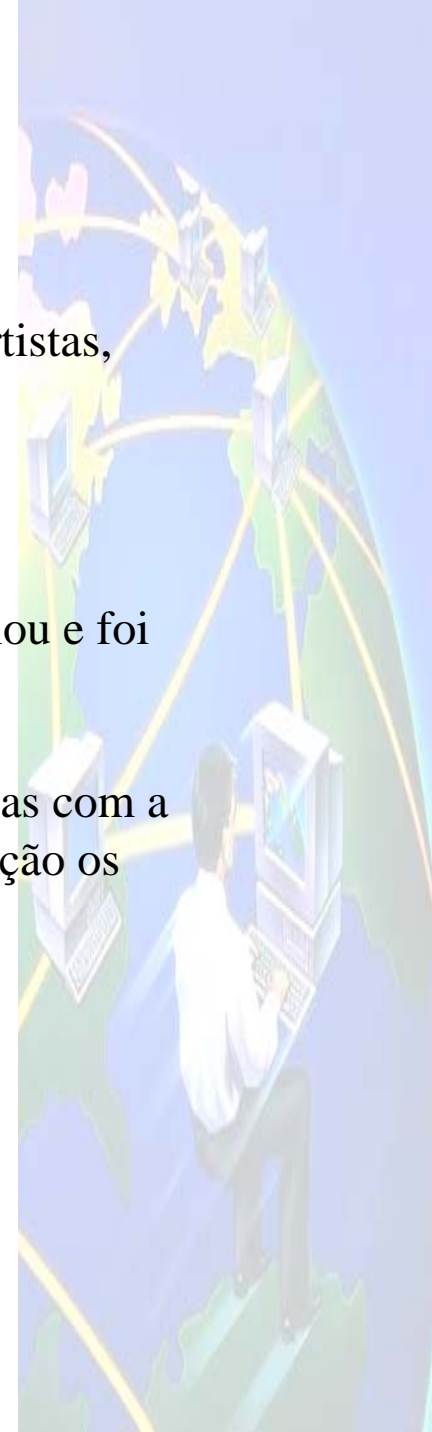
The Beatles on amazon.com

- Abbey Road 1990
- The Beatles (The White Album) 1990
- Sgt. Pepper's Lonely Hearts Club Band 1990
- Rubber Soul 1990
- Revolver [UK] 1990

Network diagram nodes: The Beatles (center), Led Zeppelin, The Rolling Stones, Bob Dylan, Johnny Cash, Tom Petty, Robert Plant, Roy Orbison, Traveling Wilburys, Paul McCartney, John Lennon, Ringo Starr, George Harrison, Eric Clapton, Jimi Hendrix.



- É um enorme banco de dados com informações sobre vários artistas, bandas e filmes
- Funciona de forma simples
- O *site* monta um mapa de uma rede que mostra quem influenciou e foi influenciado pela sua banda ou artista escolhido para pesquisa
- Mapeia as pesquisas de música e filmes através da Amazon, mas com a vantagem de nos mostrar de forma gráfica e de fácil memorização os resultados das pesquisas
- Inova também, por trabalhar com similaridade nos resultados





⚠ O Texto para Análise é obrigatório.

Análise textual (até 15.000 caracteres):

Fontes: *(Ctrl ou Shift para múltipla escolha)*

- TODAS
- Agência Brasil
- Agência ComprasNet
- Agência CT
- Agência Câmara

Período: até (dd/mm/aaaa)

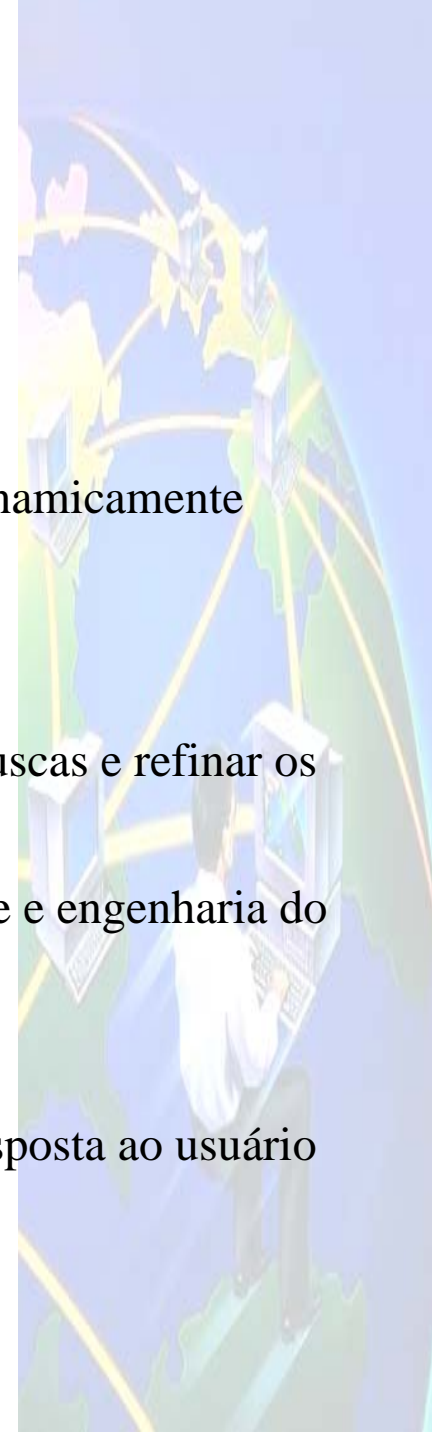
Mostrar Gráfico nos Resultados

ANALISAR TEXTO

LIMPAR

VOLTAR

- Última geração em tecnologia de ferramentas de busca
- Utiliza tecnologias inteligentes (IA)
 - PCE - Pesquisa Contextual Estruturada
 - RC2D - Representação do Conhecimento Contextualizado Dinamicamente
 - Mineração de Dados
 - Raciocínio Baseado em Casos
 - Engenharia de Ontologias
- Utiliza semânticas e estruturas valorativas para contextualizar as buscas e refinar os resultados
- Hierarquização de conteúdos com base em métricas de similaridade e engenharia do conhecimento
- Análises históricas e buscas com grandes volumes de texto
- Respostas qualitativas e documentos efetivamente relevantes na resposta ao usuário
- Permite resultados em gráficos acoplados a textos
- A Base de Dados ainda é temática focada em Governo Eletrônico



Referências

BLATTMANN, Ursula, **FACHIN**, Gleisy R. B, **RADOS**, Gregório J. Varvakis. *Recuperar a informação eletrônica pela Internet, disponível na Internet via URL: www.ced.ufsc.br/~ursula/papers/buscanet.html*. (18/04/2006)

CENDON, Beatriz Valadares. *Ferramentas de busca na Web, disponível na internet via URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100006*. (18/04/2006)

Ciência da Informação – Web search tools - <http://www.scielo.br>

How Internet Search Engines Work - <http://computer.howstuffworks.com/search-engine2.htm>. (25/04/2006)

<http://searchenginewatch.com> (25/04/2006)

BAEZA - YATES, R. Baeza-Yates, B., Ribeiro-Neto, *Modern Information Retrieval*. New York: ACM Press Series/Addison Wesley, 1999.

KAUFMANN, Morgan, Chakrabarti, Soumen. *Mining the Web: Discovering Knowledge from Hypertext Data.*,2002.

Universidade do Amazonas – Dpto Ciências da Computação – Tecnologia Web em coletas de páginas - <http://www.dcc.ufam.edu.br/~alti/tw/slides/TW-Aula%204%20-%20Crawling%20-%20Alunos.pdf> (18/04/2006)

Sistemas de busca da web: diretórios e mecanismos de busca - http://www.quatrocantos.com/tec_web/sist_busca/sb_sum.htm (18/04/2006)

