



Recuperação de Informação: Apresentação da Disciplina e Conceitos Básicos

Prof. Luciano Barbosa

(Material adaptado dos slides dos livros adotados)



A Disciplina

- Objetivos:
 - Aprender principais técnicas da área de RI
 - E como engenhos de busca são construídos



Bibliografia

- Information Retrieval in Practice. B. Croft, D. Metzler, T. Strohman. Pearson Education, 2009.
- Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. Schuumltze. Cambridge UP, 2008.
- Modern Information Retrieval. R. Baeza-Yates, B. Ribeiro-Neto. ACM Press Books, 2011.



Projeto: Construção de um Engenho de Busca Vertical



Tudo Notícias Artigos Jurisprudência Diários Legislação Modelos e peças

Página 1 de 1.771.005 resultados para "eleição"



Eleições (Sinônimo de **Eleição**)

Ato pelo qual o povo escolhe mediante sufrágio ou aclamação uma ou mais pessoas que recebem a delegação de representá-lo, exercendo determinada função.

Tópico • 31 seguidores

2014 - Eleições e Copa do Mundo.

que as **eleições** gerais, que será realizada logo após a Copa do Mundo, sejam decididas não pelos quase 142 milhões... para a **eleição** presidencial do país e de alguns governos estaduais, a exemplo de Mato Grosso. Nós

Artigo • Antonio • 04/06/2014


Voto nulo e novas eleições

De dois em dois anos, em **eleições** municipais ou regionais, sempre surge alguém para hastear a necessidade de marcação de nova **eleição** se a nulidade atingir mais de metade dos votos do país... decorre da constatação de fraude nas ...

Artigo • Danielli • 11/09/2014



Projeto: Construção de um Engenho de Busca Vertical

 Healthline

Topics & Tools ▼

q

Results for **coconut oil**

Showing 1 to 10 of 44 results

Everything

Articles

Blogs

Interactive Tools

Coconut Oil Diet: Weight Loss Fact or Fiction?

Talk of **coconut oil** being the latest weight loss sensation is pretty widespread these days. But is this claim grounded in fact or just a bunch of hype? "Some people claim that **coconut oil** increases energy, improves heart health, and helps with weight...

Can I Use Coconut Oil for Skin Care?

Coconut oil has been used to fight dry skin for centuries. It is often recommended to treat chronic dry skin, eczema, psoriasis, and it's also used as an oil massage for infants. It is commonly applied to the skin after a bath or shower to help the skin...

Can I Use Coconut Oil for Hair Growth?

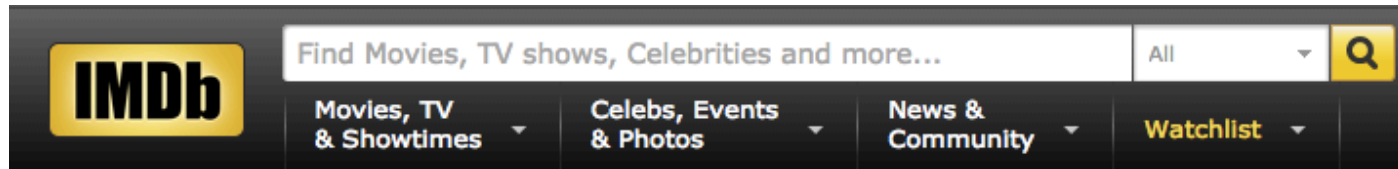
Coconut oil is produced when the coconut meat is removed from the outer hard shell and pressed. Lately, **coconut oil** is being touted as a panacea for all sorts of ailments, from indigestion to asthma to autism. Now, some are suggesting a link between coconut...

The Health Benefits of Coconut Oil

Coconut oil is derived from the white "meat" of mature coconuts. Coconut oil is extracted by a press that separates the oil from the fruit itself.



Projeto: Construção de um Engenho de Busca Vertical




Most Popular Titles With Quotes Matching "violence"

1-50 of 998 titles.

1. [Game of Thrones](#) (2011 TV Series)
Episode: [A Man Without Honor](#) (2012)
Have you gone soft, Clegane? I always thought you had a talent for violence. Burn the villages, burn the farms. Let them know what it means to choose the wrong side.
2. [Game of Thrones](#) (2011 TV Series)
Episode: [The Broken Man](#) (2016)
Violence is a disease. You don't cure a disease by spreading it to more people.
3. [Game of Thrones](#) (2011 TV Series)
Episode: [A Golden Crown](#) (2011)
Where do I begin, my lords and ladies? I am a vile man, I confess it. My crimes and sins are beyond counting. I have lied and cheated, gambled and whored. I'm not particularly good at violence, but I'm good at convincing others to do violence for me. You want specifics, I suppose. When I was seven, I saw a servant girl bathing in the river. I stole her robe and she was forced to return to the castle naked and in tears. I close my eyes, but I can still see her tits bouncing...
4. [Game of Thrones](#) (2011 TV Series)
Episode: [No One](#) (2016)
I choose violence.



Projeto: Construção de um Engenho de Busca Vertical



Faça **login** ou **cadastre-se**


PedidosMinhas ListasÚltimas Compras


Buscar


Minha Cesta
R\$ 0,00
Cesta Vazia


0


Entregar em: [Escolha o endereço](#)


Telefonia


Informática

Eletrodomésticos

Eletroportáteis

Alimentos

Serviços

Veja todos os departamentos

Categoria

— Higiene Bucal (64)

Preço

— R\$0 - R\$10 (51)

— R\$10 - R\$25 (12)

Home / busca: **creme dental**

1 - 12 de 64 Produtos encontrados para **creme dental**

Visualização: **Grade** | Lista

< 1 2 3 4 >

Ordenar por: Popularidade



Creme Dental COLGATE Tripla Ação Hortelã 90g

R\$ 2,59



Creme Dental COLGATE Total 12 90g Cada Leve 4 Pague 3

R\$ 17,37
LEVE 1 por R\$ 13,90 unid
ou
LEVE 2 por R\$ 13,03 unid
ou
LEVE 3 por R\$ 11,55 unid



Pack Creme Dental CLOSEUP Diamond Attraction Leve 3 Pague 2

R\$ 9,38



Creme Dental COLGATE Tripla Ação 90g

R\$ 2,59





Projeto e Avaliação

- Projeto: engenho de busca vertical
 - Busca em páginas em um mesmo domínio
 - Página com “estrutura”
 - Ex:
 - Produtos (câmera, alimentos, carros etc)
 - Filmes, músicas, emprego etc
 - Podem ser usadas APIs online
- Avaliação (modelo de empresa)
 - Treinamento
 - Acompanhamento
 - Demo



Primeiro Módulo

- Introdução ao Curso
- Funcionamento de Engenhos de Busca
- Coleta de Dados
- Processamento de Texto
- Classificação de Texto
- Extração de Dados na Web



Segundo Módulo

- Indexação de Dados
- Busca:
 - Modelos de RI
 - Ranking e Processamento de Consultas
 - Refinamento de Consultas
 - Avaliação de Sistemas de RI
 - Análise de Links



Recuperação de Informação: Origens

- Buscar documentos em bibliotecas

| Field to search | Enter Word(s) | Word(s) as Phrase? | Results | Total |
|---|----------------------|---|---------|-------|
| <input checked="" type="checkbox"/> Keyword | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Author | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Corporate Author | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Title | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Journal Title | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Subject | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Genre | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Series | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Table of Contents | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Notes | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Publisher | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> ISBN | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> ISSN | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> Call Number (LC) | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |
| <input type="checkbox"/> System Number | <input type="text"/> | <input checked="" type="radio"/> No <input type="radio"/> Yes | | |

Search by:

Year from Year to

Location

Limit by format

Language

[Search](#) | [Reset form](#)



Recuperação de Informação

- Lida com representação, armazenamento e acesso a itens de informação:
 - Documentos, páginas Web, dados estruturados, objetos multimídia
- Estruturas de dados para rápido acesso: índice

| | |
|------|--|
| QA76 | Clark, Michael |
| .55 | Cultural treasures of the Internet / M. Clark. -- Upper |
| .C58 | Saddle River, N.J. : Prentice Hall, 1995. |
| 1995 | xxix, 313 p. : ill. ; 24 cm. |
| | ISBN 0132096692. |
| | 1. Computers--Cultural Impact. 2. Internet--Cultural Impact. |



Objetivo de RI

- Satisfazer a necessidade de informação do usuário
- Exemplo:
 - “Encontrar todos os documentos sobre o Governo Federal relacionados ao financiamento dos estádios da Copa do Mundo”
- Não necessariamente uma boa estrutura de consulta
- Usuário precisa “traduzir” para um conjunto de termos
- Recuperar todos documentos **relevantes** e o menor número de **não-relevantes**



O que é um Documento?

- Exemplos: páginas Web, livros, notícias, artigos, mensagens de textos, posts, pdfs etc
- Propriedades de um documento:
 - Texto
 - Alguma estrutura



Estrutura em Documentos

- Artigos: título, autor, instituição etc

Efficiently Linking Text Documents with Relevant Structured Information

Venkatesan T. Chakaravarthy Himanshu Gupta Prasan Roy Mukesh Mohania
IBM India Research Lab
New Delhi, India
{vechakra, higupta3, prasanr, mkmukesh}@in.ibm.com

ABSTRACT

Faced with growing knowledge management needs, enterprises are increasingly realizing the importance of interlinking critical business information distributed across structured and unstructured data sources. We present a novel system, called EROCS, for linking a given text document with relevant structured data. EROCS views the structured data as a predefined set of “entities” and identifies the entities that best match the given document. EROCS also embeds the identified entities in the document, effectively creating links between the structured data and segments within the document. Unlike prior approaches, EROCS identifies such links even when the relevant entity is not explicitly mentioned in the document. EROCS uses an efficient algorithm that performs this task keeping the amount of information retrieved from the database at a minimum. Our evaluation shows that EROCS achieves high accuracy with reasonable overheads.

structured data consists of all information about sales transactions, customers and products. The organization, with a network of multiple stores, has a steady inflow of complaints into a centralized complaint repository; these complaints are accepted using alternative means, such as a web-form, email, fax and voice-mail (which is then transcribed). Each such complaint is typically a free-flow narrative text about one or more sales transactions, and is not guaranteed to contain the respective transaction identifiers; instead, it might divulge, by way of context, limited information such as the store name, a partial list of items bought, the purchase dates, etc. Using this limited information, EROCS discovers the potential matches with the transactions present in the sales transactions database and links the given complaint with the matching transactions.

Such linkage provides *actionable* context to a typically fuzzy, free flow narrative which can be profitably exploited in a variety of ways.



Documentos vs. Entradas em BD: Estrutura

- Campos (ou atributos) bem definidos
- Tuplas em BD relacionais
- Semântica bem definida

| _id | GRUPOS... | GRUPOS... | SERVICO... | SERVICO... | LOGRAD... | NUMERO | BAIRRO | RPA | DATA_DE... | SITUACA... |
|-----|-----------|------------|------------|-------------|-------------|--------|------------|-----|-------------|------------|
| 1 | 7 | ARBORIZ... | 8 | PODA DE... | AV BEIRA... | 00 | TORRE | 4 | 2016-07-... | ATENDIDA |
| 2 | 7 | ARBORIZ... | 8 | PODA DE... | AV MAUR... | 00 | IPUTINGA | 4 | 2016-07-... | ATENDIDA |
| 3 | 7 | ARBORIZ... | 16 | VISTORIA... | PRC PRO... | 00 | JAQUEIRA | 3 | 2016-07-... | CADAST... |
| 4 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA CAR... | 381 | HIPODR... | 2 | 2016-07-... | CADAST... |
| 5 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA CAS... | 00 | MADALENA | 4 | 2016-07-... | CADAST... |
| 6 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA CELIA | 00 | MUSTAR... | 5 | 2016-07-... | CADAST... |
| 7 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA ES... | 255 | GRACAS | 3 | 2016-07-... | CADAST... |
| 8 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA FLO... | 0 | SAO JOSE | 1 | 2016-07-... | CADAST... |
| 9 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA FRA... | 378 | SANTO A... | 1 | 2016-07-... | CADAST... |
| 10 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA GAR... | 275 | CAMPIN... | 2 | 2016-07-... | CADAST... |
| 11 | 7 | ARBORIZ... | 16 | VISTORIA... | RUA HA... | 51 | CAMPO ... | 2 | 2016-07-... | CADAST... |



Documentos vs. Entradas de BD: Consultas

- BD:
 - Casa com os campos no BD
 - “Select * From Acidentes WHERE bairro=‘Ipatinga’ AND data=‘2016’”
- RI:
 - Procura em todo o texto pela ocorrência das palavras
 - “Acidentes em Ipatinga em 2016”
 - Casamento exato de palavras não é suficiente
 - Ex.: “uma batida de carro ocorreu na rua ...”



Várias Dimensões de RI

| Content | Applications | Tasks |
|--------------|-------------------|--------------------|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | P2P search | |
| | Literature search | |



Tarefas de RI

- Busca ad-hoc
 - Encontrar documentos relevantes para uma consulta arbitrária
- Filtragem
 - Identificar documentos relevantes baseado em perfil de usuários
- Classificação
 - Encontrar rótulos para documentos
- Pergunta e resposta
 - Dar uma resposta específica a uma pergunta



A Era da Web

- RI até recentemente restrita a bibliotecários
- A Web mudou isso
 - Autoria descentralizada e fácil acesso
 - Maior repositório de conhecimento na história da humanidade



Web Mudou a Busca por Informação

- Busca na Web é hoje a aplicação mais importante de RI
- Desafios:
 - Grande volume de dados: mais de 20 bilhões de páginas
 - Grande volume de consultas: Google 40 mil/s
- Desempenho e escalabilidade são cruciais!



Desafios para Coleta

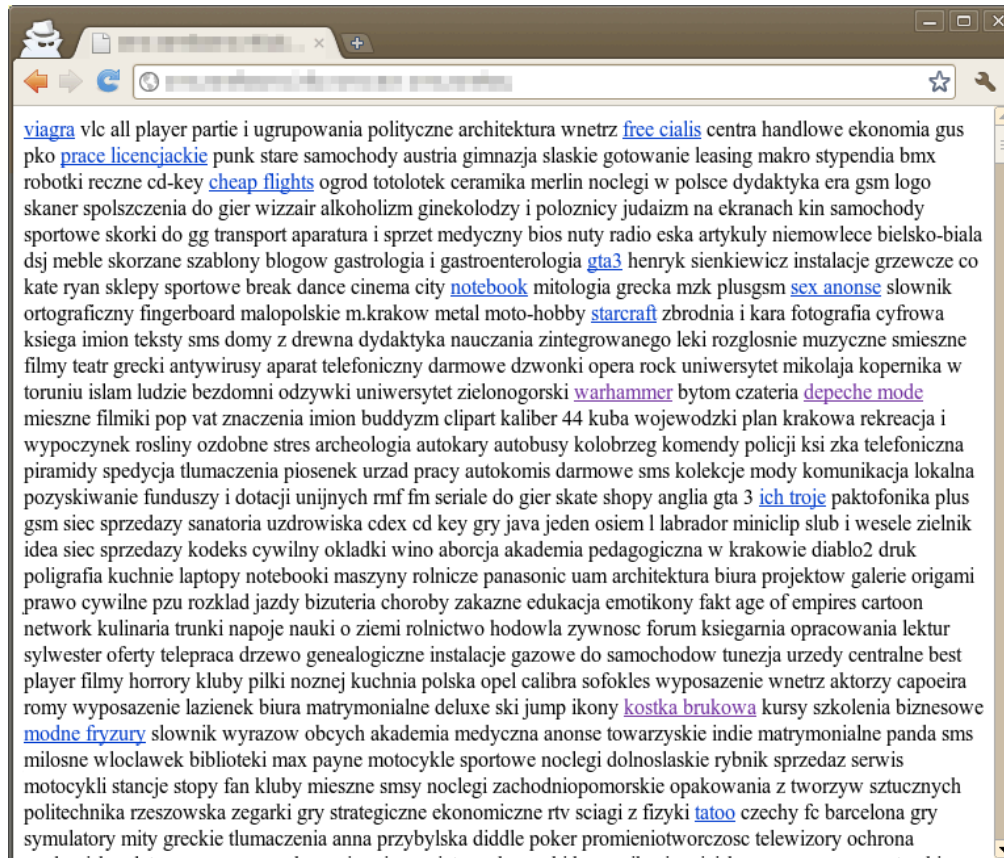
- Dados distribuídos: milhões de sites ativos
- Dinamicidade: 80% desapareceu em 1 ano [Dasgupta et al., 2007]
- Documentos em diferentes formatos





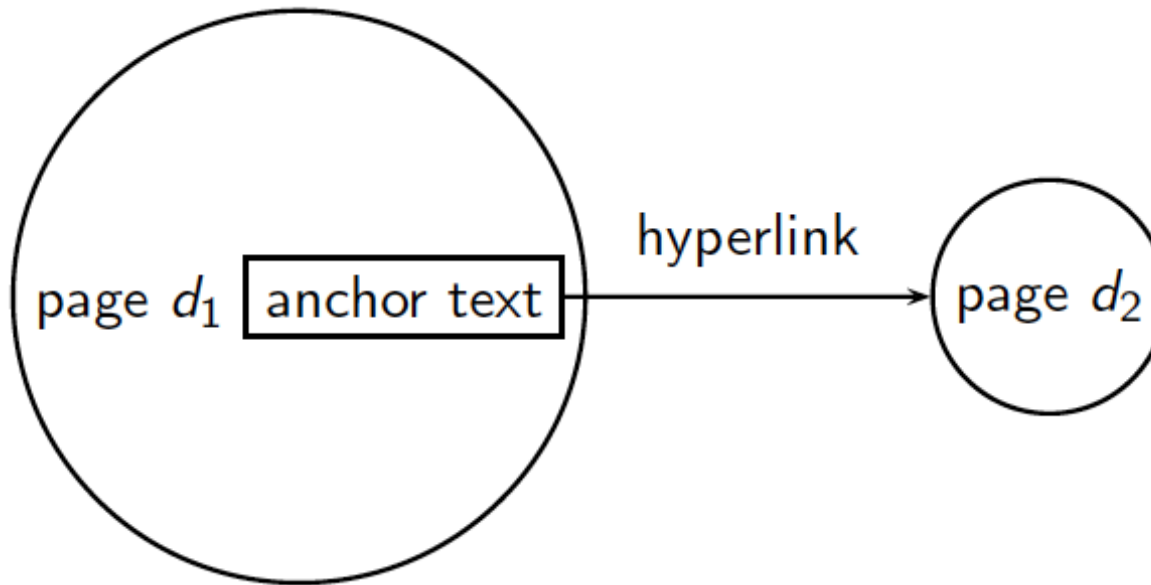
Outro Desafio: Spam

- Texto não informativo criado para dar visibilidade à página
- Afeta a eficiência e qualidade dos resultados





A Web é um Grafo Direcionado





RI e Engenheiros de Busca

Information Retrieval

Relevance

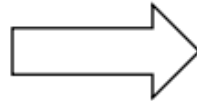
- Effective ranking*

Evaluation

- Testing and measuring*

Information needs

- User interaction*



Search Engines

Performance

- Efficient search and indexing*

Incorporating new data

- Coverage and freshness*

Scalability

- Growing with data and users*

Adaptability

- Tuning for applications*

Specific problems

- e.g. Spam*



Yahoo!

- Iniciou em 1994 como diretório

computer chess - Yahoo! Search Results

http://us.search.yahoo.com/search?fr=io&p=computer+chess

computer chess

Getting Started Soon... Later... Fun News MMedia CS Funds Students Conf CriThi Admin HOME BioBDx TODAY webcast Dashboard - Google... Most Visited

Google "computer chess" Search Options Customize

Web Images Video Local Shopping more

computer chess

1 - 10 of 29,700,000 for computer chess (About) - 0.02 s | SearchScan

Also try: [play free computer chess](#), [computer chess games](#), [More...](#)

Computer Chess Games
Get cashback on Computer Games. Search Now and Save.
[Search.Live.com/cashback](#)

Play Free Chess Against Computer
Find 1000's of items and Compare prices at Smarter.com.
[www.smarter.com](#)

Computer chess - Wikipedia, the free encyclopedia
1990s Pressure-sensory **Chess Computer** with LCD screen ... Since then, **chess** enthusiasts and **computer** engineers have built, with increasing ...
[en.wikipedia.org/wiki/Computer_chess](#) - 131k - [Cached](#)

Chess engine - Wikipedia, the free encyclopedia
(Redirected from **Chess computer**) Jump to: navigation, search ... 1992 became the first microcomputer to win the World **Computer Chess** Championship ...
[en.wikipedia.org/wiki/Chess_computer](#) - [Cached](#)

Play Chess Against the Computer - Chess.com
Live **Chess**. Against the **Computer**. Vote **Chess**. Learn. **Chess** Rules & Basics. Openings - Game Explorer ...
[Chess.com](#) Gear. Classified Ads. Members. Search Members ...
[www.chess.com/play/computer.html](#) - [Cached](#)

Thinking Machine 4: Play the Game
About | Image gallery | Home | Contact ...
[turbulence.org/spotlight/thinking/chess.html](#) - [Cached](#)

Home - Computer Chess Wiki
... trusted source for **computer chess** information! Table of ... If you are a **computer chess** expert and want to write a new article, then put it up on our site!

Sponsor Results

The Chess Store - Chess Computers
All Excalibur and Saitek models. Live service, no tax, free gift wrap.
[www.thechessstore.com](#)

Marble Chess Sets
All items 30% off. Brand name **chess** sets, chessmen, boards and tables ...
[www.gammonvillage.com](#)

Chess 3D - Software - \$9.95
ProSoft **Chess** 3D is a great **Chess** Game for your **computer**. Download.
[www.prosoft3d.com](#)

Free 3D Arcade Chess
Play **Chess** Anytime w/ this Stunning 3d Arcade **Chess** Game Free.
[Playtoad.com/Chess](#)

Computer Checkers
Help Keep Your Family Protected Online. Get Special Offers.
[www.Microsoft.com/Windows](#)

Chess Computers Wholesale
Chess computers & handhelds from Saitek & Excalibur at Wholesale.



Lycos!

http://search.lycos.com/?query=computer+chess&x=0&y=0


computer chess

Getting Started Soon... Later... Fun News MMedia CS Funds Students Conf CriThi Admin HOME BioBDx TODAY webcast Dashboard - Google... Most Visited

Google northernlight Search Bookmarks PageRank Check AutoLink AutoFill Send to Settings

Lycos Home Lycos Mail

WEB IMAGE VIDEO PEOPLE BUSINESS NEWS

 **LYCOS** computer chess GO GET IT! Advanced Search

Also Fetch!: free computer chess games, free online computer chess, play computer chess, More...

Computer Chess Games - Search.Live.com/cashback
Get Cashback on computer chess games. Search, Shop and Save.

Play Free Chess Against Computer - www.smarter.com
Find 1000's of items and Compare prices at Smarter.com.

The Chess Store - Chess Computers - www.thechessstore.com
All Excalibur and Saitek models. Live service, no tax, free gift wrap.

Marble Chess Sets - www.gammonvillage.com
All items 30% off. Brand name chess sets, chessmen, boards and tables - while supplies last.

SPONSORED RESULTS

Chess 3D - Software - \$9.95
ProSoft Chess 3D is a great Chess Game for your computer. Download.
www.prosoft3d.com

Free 3D Arcade Chess
Play Chess Anytime w/ this Stunning 3d Arcade Chess Game Free.
Playtoad.com/Chess

Computer Checkers
Help Keep Your Family Protected Online. Get Special Offers.
www.Microsoft.com/Windows

Chess Computers Wholesale
Chess computers & handhelds from Saitek & Excalibur at Wholesale.
www.WholealeChess.com

Chess Computers - 40% Off Sale
Huge Selection of Computer Chess on Sale. Low Prices & Fast Shipping.
www.ChessSets.com

Fatal1Ty: I'm a PC
Find Out How Fatal1ty Perfects His Gaming w/ a PC. Tell Us Your Story.
ImAPC.LifeWithoutWalls.com

Web Results 1 thru 10 of 832,101 (Info)

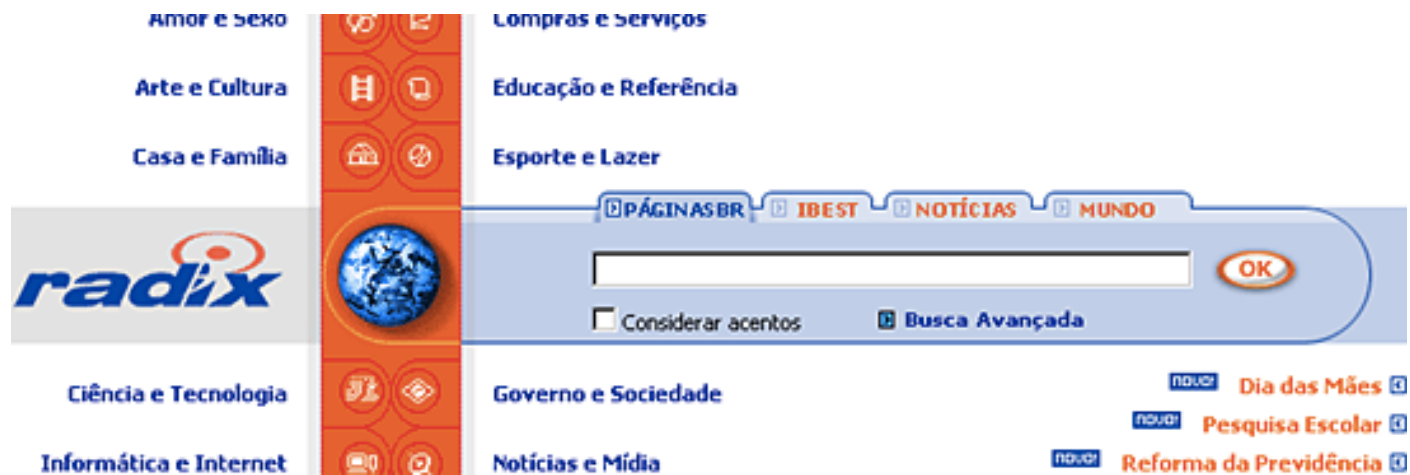
Computer chess - Wikipedia, the free encyclopedia
1990s Pressure-sensory Chess Computer with LCD screen ... Since then, chess enthusiasts and computer engineers have built, with increasing ...
en.wikipedia.org

Play Chess Against the Computer - Chess.com
Live Chess. Against the Computer. Vote Chess. Learn. Chess Rules & Basics. Openings - Game Explorer ... Chess.com Gear. Classified Ads. Members. Search Members ...
www.chess.com

Thinking Machine 4: Play the Game
About | Image gallery | Home | Contact ...
turbulence.com



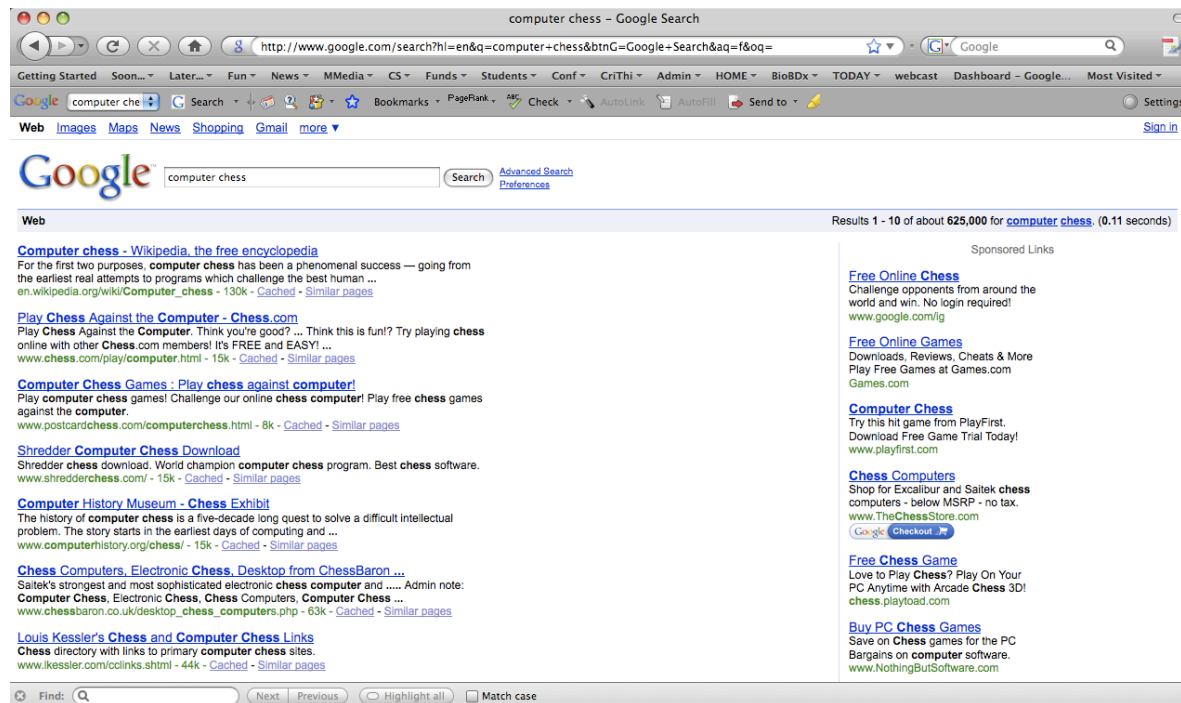
Radix





Google

- Criado no começo dos anos 2000 por estudantes de doutorado
- Inovou com o algoritmo de PageRank





Baidu

- Engenho de busca mais popular da China

🎤📷百度一下



Quora

- Especializado em Q&A em linguagem natural

Quora [Ask Question](#) [Read](#) [Answer](#) [Notifications](#) ² [Luciano](#)

Feeds [Edit](#)

Top Stories

Bookmarked Answers

- Movies
- Visiting and Travel
- Sports
- Economics
- Education


Trending Now


- Stranger Things
- Didi Chuxing-Uber's China Operations Acquisition
- Khizr Khan's Speech at the 2016 DNC
- KickAss Torrents Owner Arrested, Faces Extradition to US

Top Stories For You

Popular on Quora

What is the biggest mistake that a big company has made?

 **Andy Dowling**
472.4k Views



McDonald's and the 1984 Olympics. The 1984 Summer Olympics were hosted in Los Angeles, California. McDonald's was a huge sponsor of the games, and decided to use the games as a way to advertise. The... [\(more\)](#)

[Upvote](#) **17.3k** [Downvote](#) [Comments](#) **42+**

[f](#) [t](#) [s](#) [...](#)

Improve Your Feed


- ✓ [Visit your feed](#)
- ✓ [Follow 10 more topics](#)
- ✓ [Find your friends on Quora](#)
- ✓ [Upvote 5 more good answers](#)
- ✓ [Ask your first question](#)
- ✓ [Add info about what you know](#)
- ✓ [Answer a question](#)



[Sessions - Now Taking Questions](#) [View All](#) >









WolframAlpha

- Busca sobre bases de conhecimento

 **WolframAlpha** computational... knowledge engine

albert einstein  

     Examples  Random

Assuming "albert einstein" is a person | Use as [a public school](#) or [an artwork](#) instead

Input interpretation:

Albert Einstein (physicist)

Basic information:

| | |
|----------------|--|
| full name | Albert Einstein |
| date of birth | Friday, March 14, 1879 (137 years ago) |
| place of birth | Ulm, Baden-Wurttemberg, Germany |
| date of death | Monday, April 18, 1955 (age: 76 years) (61 years ago) |
| place of death | Princeton, New Jersey, United States |