University of Twente

Information Retrieval Modeling

Russian Summer School in Information Retrieval

Djoerd Hiemstra
http://www.cs.utwente.nl/~hiemstra

University of Twente
The Netherlands

PART 1
the basics

Goal

- Gain basic knowledge of IR
- Intuitive understanding of difficulty of the problem
- Insight in consequences of modeling assumptions
- *biased* comparison of formal models

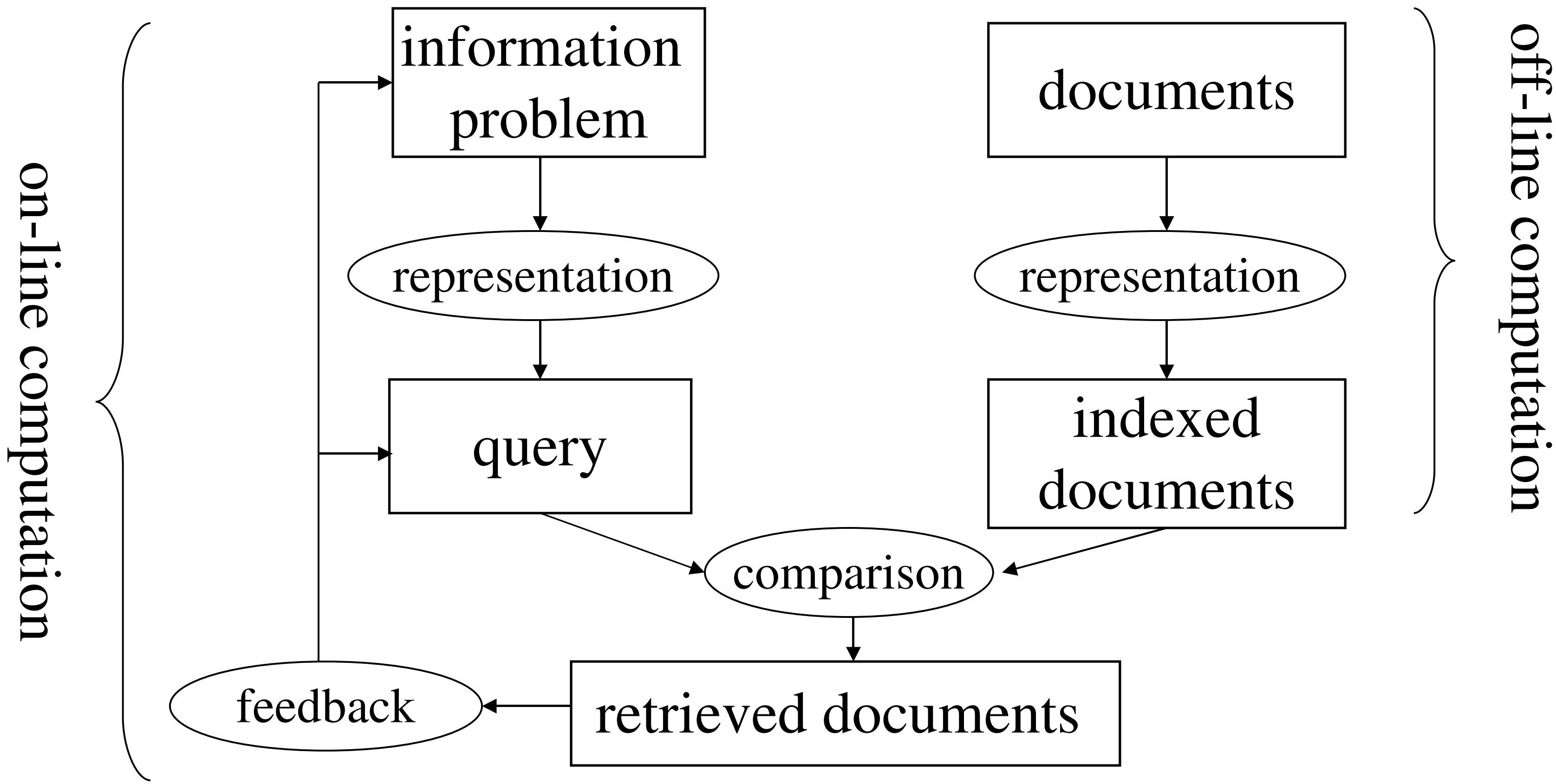
Overview

1. Boolean retrieval
2. Vector space models
3. Probabilistic retrieval / Naive Bayes
4. Google's PageRank
5. The QUIZ

Course material

- Djoerd Hiemstra, “Information Retrieval Models”, In: Ayse Goker, John Davies, and Margaret Graham (eds.), *Information Retrieval: Searching in the 21st Century, Wiley*, 2009.

Information Retrieval



Full text information retrieval

- Index based on uncontrolled (free) terms (as opposed to controlled terms)
- Every word in a document is a potential index term
- Terms may be linked to specific XML elements in a text (title, abstract, preface, image caption, etc.)

Full text information retrieval

- Different views on documents
  - External: data not necessarily contained in the document (metadata)
  - Logical: e.g. chapters, sections, abstract
  - Layout: e.g. two columns, A4 paper, Times
  - Content: the text

this is what IR models are about

mostly...

Full text information retrieval

- Automatic processing of natural language:
  - statistics (counting words)
  - stop list
  - morphological stemming
  - part-of-speech tagging
  - compound splitting
  - partial parsing: noun phrase extraction
  - other: use of thesaurus, named entity recognition, ...

this is what IR models are about

mostly...

Full text information retrieval

- stop list
  - remove frequent words (the, and, for, etc.)
- stemmer
  - rewrite rules, rules of the thumb
  - sky skies ski skiing → ski
- compound words
  - word contains more than one morpheme
  - Fietsbandventiel → fiets, band, ventiel
- phrases
  - separate words not good predictors: New York

Being an IR model

apply big billi bodi boston brought creat decid docum dump electron employe format good govern hope industri join king live lot massachusett microsoft offic open parti peopl problem recognit revolut sauc save softwar standard state tea thumb worri

Massachusetts dumps Microsoft Office

Massachusetts The people who brought you the Boston tea party, have joined in another revolution against good King Billy's Office software. The state government has decided that all electronic documents saved and created by state employees have to use open formats .

Microsoft is clearly worried. A lot of people live in Massachusetts and that is a big thumbs up for open sauce. However, it is hoping to get around the problem by applying recognition from an industry standards body for recognition of its own formats as open standards.

Being an IR model

bitterli central clear cloudi cloudier coast cold dai east easterli edg flurri forecast frost lead moder northeast part period persist plenti risk shower sleet snow south southern southwestern sunshin todai weather wind wintri

Today's weather forecast

Clear periods leading to a moderate frost in many parts away from the east coast. The northeast will be cloudier, as will the far south, here the risk of a few snow flurries. The bitterly cold easterly wind persisting.

Plenty of sunshine around, but rather cloudy in northeast, here some wintry showers. The south also rather cloudy, perhaps sleet or snow edging into southwestern and central southern parts later in day.

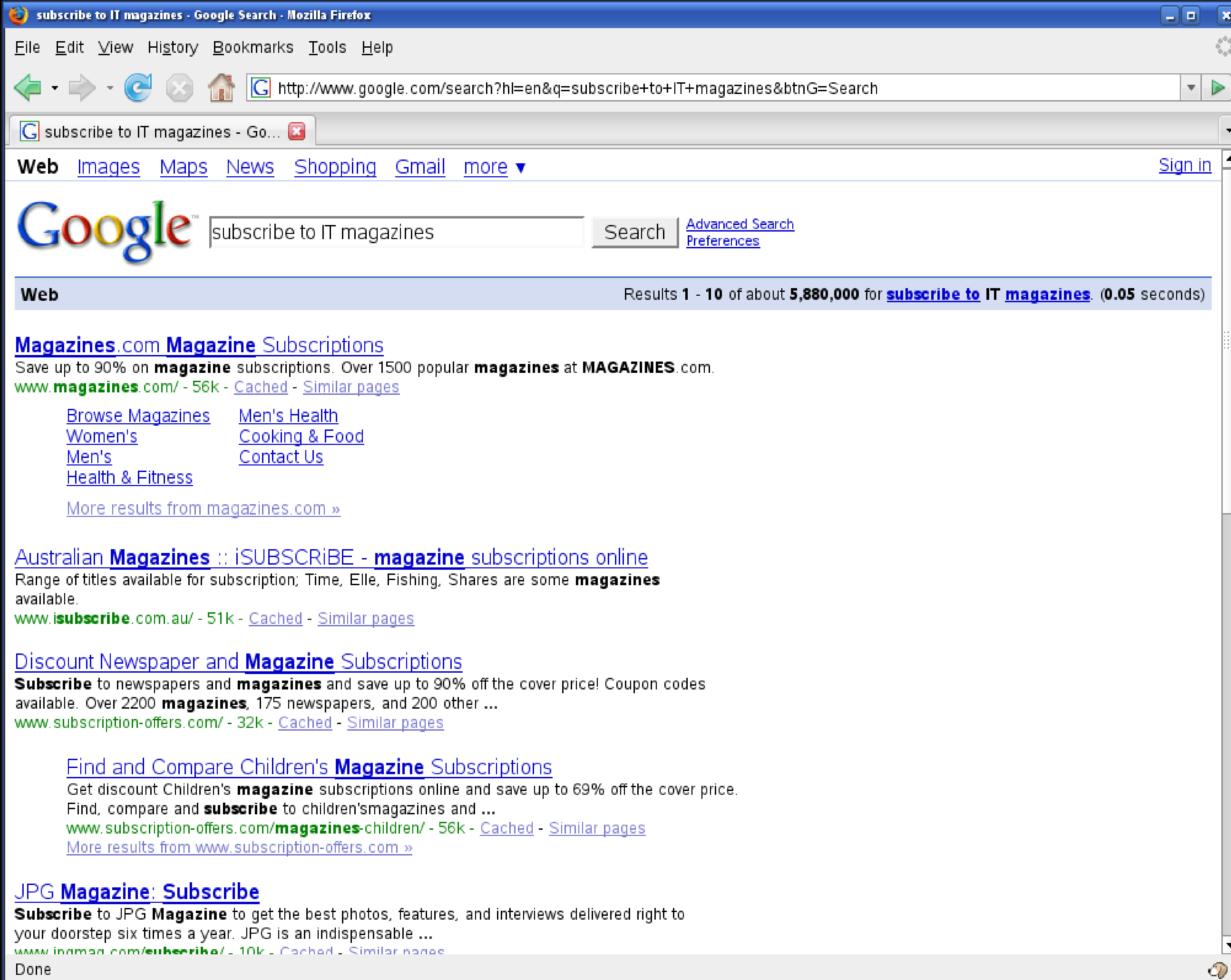
Full text information retrieval

- Advantages:
  - fully automatic indexing (saves time and money)
  - less standardisation (tailored to variation in information need of different users)
  - can still be combined (?) with aspects of controlled approach (thesaurus, metadata)

Full text information retrieval

- Main disadvantage: the (professional) user looses his/her control over the system...
  - because of 'ranking' instead of 'exact matching', the user does not *understand* why the system does what it does
  - assumptions of stop lists, stemmers, etc. do not hold universally:

    e.g. the query “last will”: are “last” or “will” stop words? should it retrieve “last would”?


subscribe to IT magazines - Google Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.google.com/search?hl=en&q=subscribe+to+IT+magazines&btnG=Search

subscribe to IT magazines - Go...

Web Images Maps News Shopping Gmail more ▾ Sign in

Google subscribe to IT magazines Search Advanced Search Preferences

Web Results **1** - **10** of about **5,880,000** for **subscribe to IT magazines**. (**0.05** seconds)

Magazines.com **Magazine** Subscriptions
Save up to 90% on **magazine** subscriptions. Over 1500 popular **magazines** at **MAGAZINES**.com.
www.**magazines**.com/ - 56k - Cached - Similar pages

Browse Magazines | Men's Health
Women's | Cooking & Food
Men's | Contact Us
Health & Fitness

More results from magazines.com »

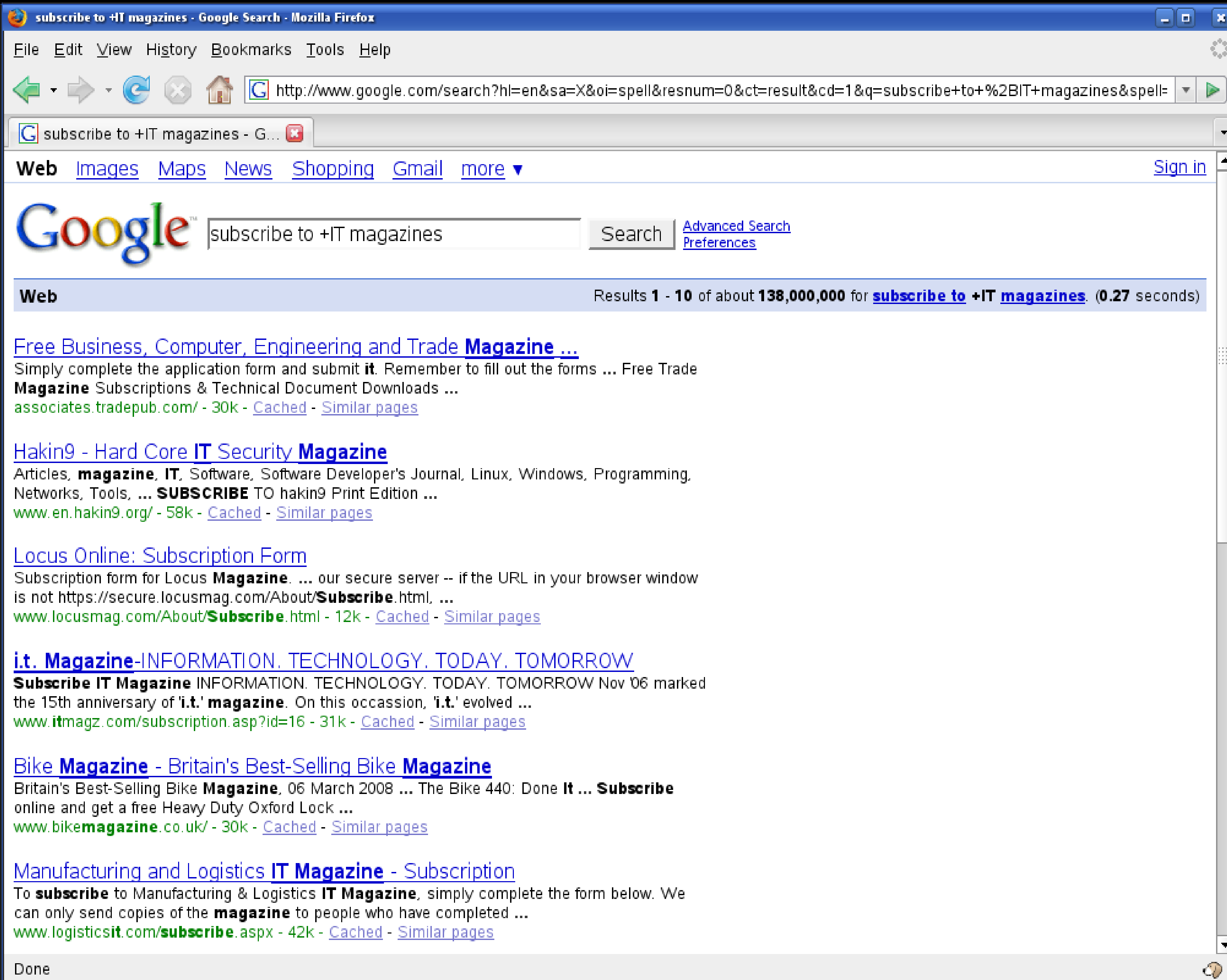
Australian **Magazines** :: iSUBSCRiBE - **magazine** subscriptions online
Range of titles available for subscription; Time, Elle, Fishing, Shares are some **magazines** available.
www.i**subscribe**.com.au/ - 51k - Cached - Similar pages

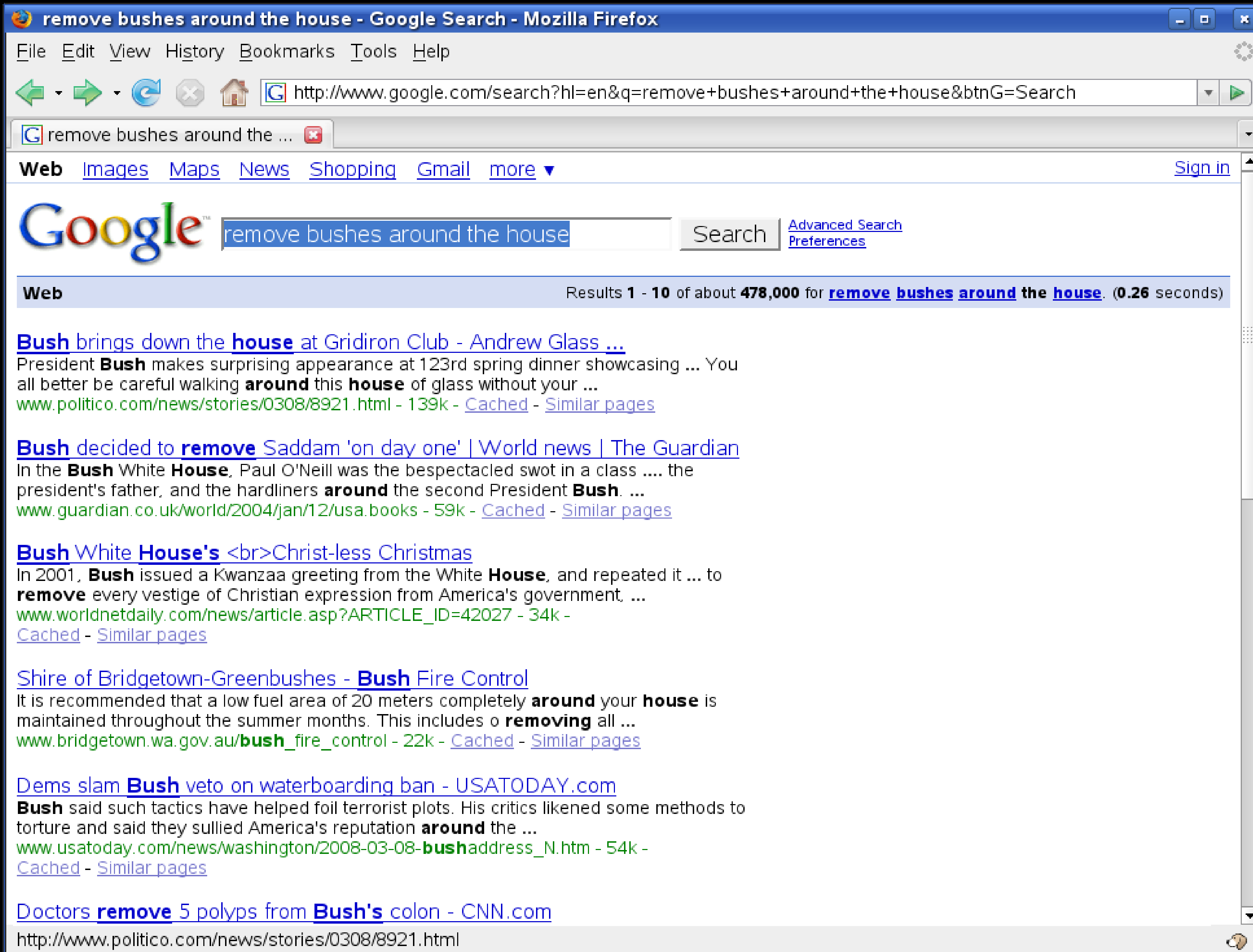
Discount Newspaper and **Magazine** Subscriptions
Subscribe to newspapers and **magazines** and save up to 90% off the cover price! Coupon codes available. Over 2200 **magazines**, 175 newspapers, and 200 other **...**
www.subscription-offers.com/ - 32k - Cached - Similar pages

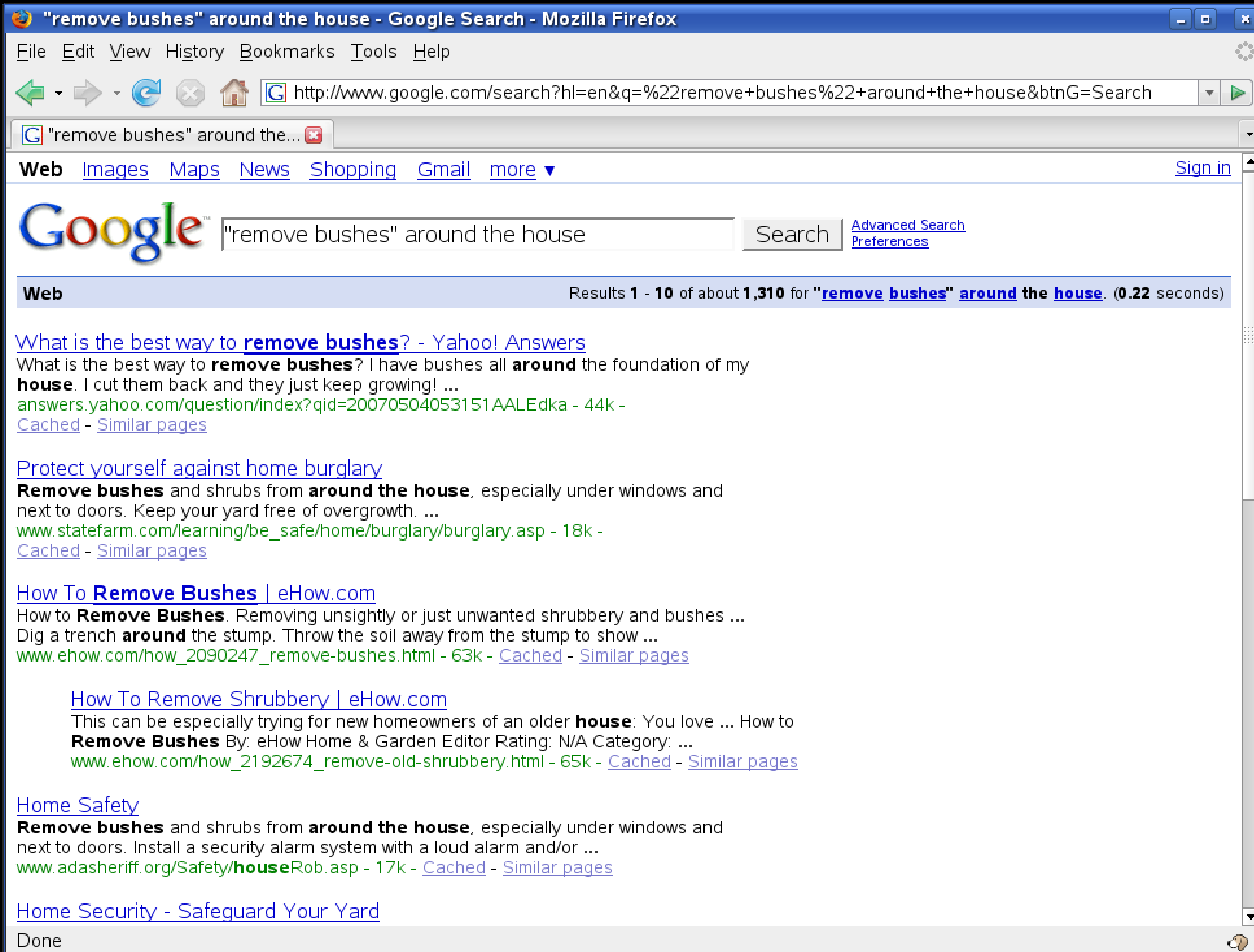
Find and Compare Children's **Magazine** Subscriptions
Get discount Children's **magazine** subscriptions online and save up to 69% off the cover price. Find, compare and **subscribe** to children'smagazines and **...**
www.subscription-offers.com/**magazines**-children/ - 56k - Cached - Similar pages
More results from www.subscription-offers.com »

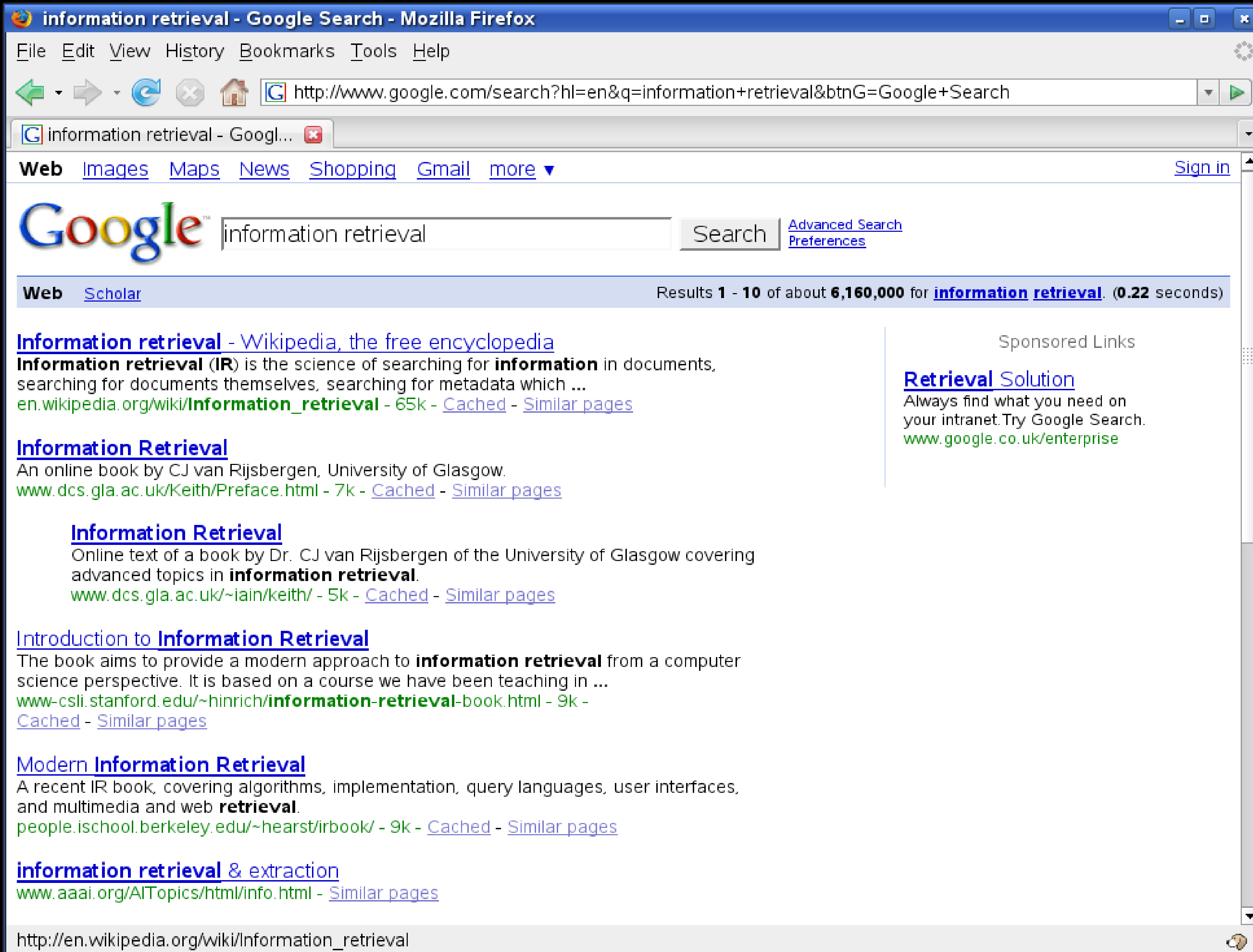
JPG **Magazine**: **Subscribe**
Subscribe to JPG **Magazine** to get the best photos, features, and interviews delivered right to your doorstep six times a year. JPG is an indispensable **...**
www.jpgmag.com/**subscribe**/ - 10k - Cached - Similar pages

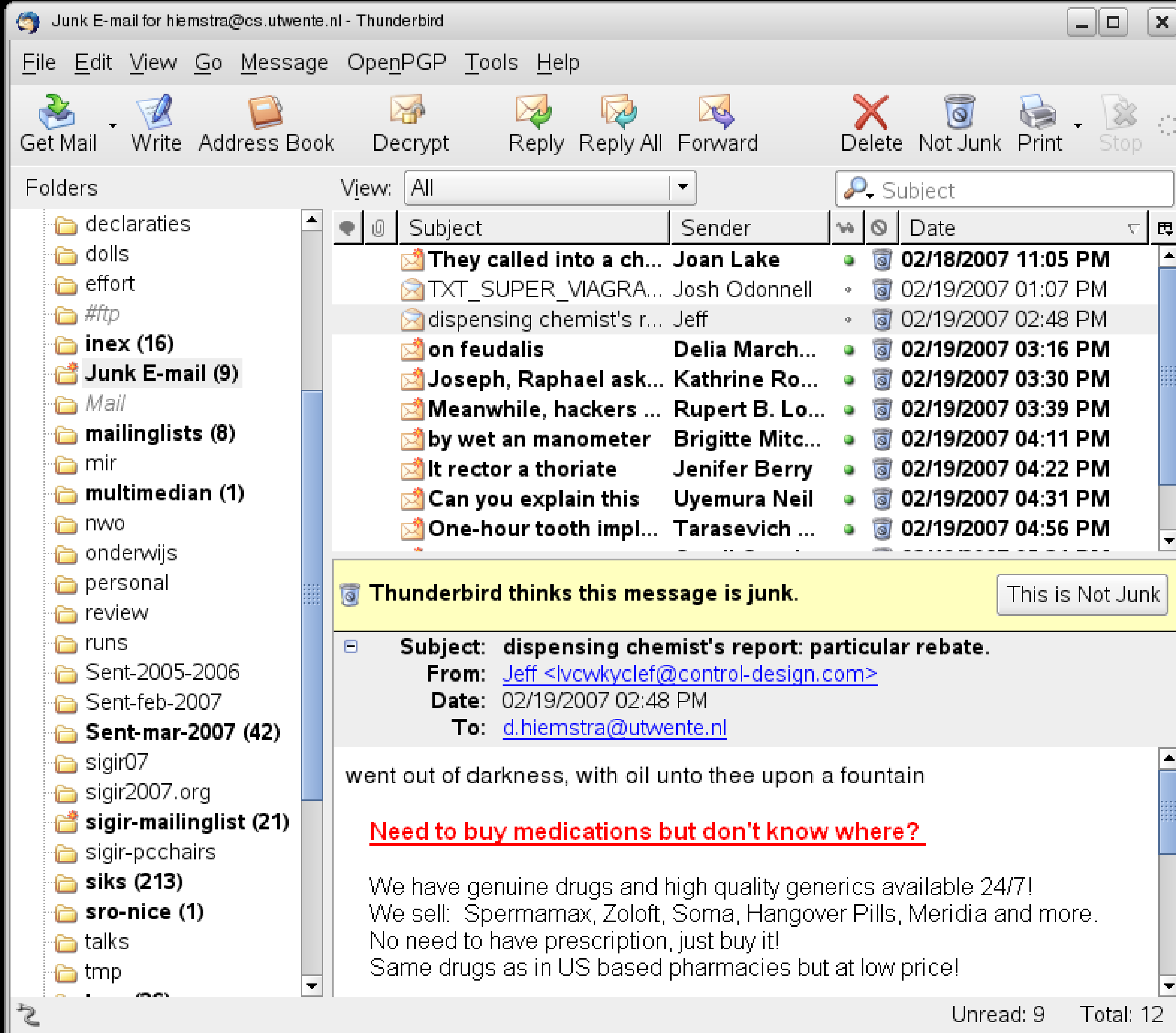
Done


subscribe to +IT magazines - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/search?hl=en&sa=X&oi=spell&resnum=0&ct=result&cd=1&q=subscribe+to+%2BIT+magazines&spell=
subscribe to +IT magazines - G...
Web Images Maps News Shopping Gmail more ▼
Sign in
subscribe to +IT magazines
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 138,000,000 for subscribe to +IT magazines. (0.27 seconds)
Free Business, Computer, Engineering and Trade Magazine ...
Simply complete the application form and submit it. Remember to fill out the forms ... Free Trade Magazine Subscriptions & Technical Document Downloads ...
associates.tradepub.com/ - 30k - Cached - Similar pages
Hakin9 - Hard Core IT Security Magazine
Articles, magazine, IT, Software, Software Developer's Journal, Linux, Windows, Programming, Networks, Tools, ... SUBSCRIBE TO hakin9 Print Edition ...
www.en.hakin9.org/ - 58k - Cached - Similar pages
Locus Online: Subscription Form
Subscription form for Locus Magazine. ... our secure server -- if the URL in your browser window is not https://secure.locusmag.com/About/Subscribe.html, ...
www.locusmag.com/About/Subscribe.html - 12k - Cached - Similar pages
i.t. Magazine-INFORMATION. TECHNOLOGY. TODAY. TOMORROW
Subscribe IT Magazine INFORMATION. TECHNOLOGY. TODAY. TOMORROW Nov '06 marked the 15th anniversary of 'i.t.' magazine. On this occassion, 'i.t.' evolved ...
www.itmagz.com/subscription.asp?id=16 - 31k - Cached - Similar pages
Bike Magazine - Britain's Best-Selling Bike Magazine
Britain's Best-Selling Bike Magazine, 06 March 2008 ... The Bike 440: Done It ... Subscribe online and get a free Heavy Duty Oxford Lock ...
www.bikemagazine.co.uk/ - 30k - Cached - Similar pages
Manufacturing and Logistics IT Magazine - Subscription
To subscribe to Manufacturing & Logistics IT Magazine, simply complete the form below. We can only send copies of the magazine to people who have completed ...
www.logisticsit.com/subscribe.aspx - 42k - Cached - Similar pages
Done


remove bushes around the house - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/search?hl=en&q=remove+bushes+around+the+house&btnG=Search
remove bushes around the ...
Web Images Maps News Shopping Gmail more ▼
Sign in
remove bushes around the house
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 478,000 for remove bushes around the house. (0.26 seconds)
Bush brings down the house at Gridiron Club - Andrew Glass ...
President Bush makes surprising appearance at 123rd spring dinner showcasing ... You all better be careful walking around this house of glass without your ...
www.politico.com/news/stories/0308/8921.html - 139k - Cached - Similar pages
Bush decided to remove Saddam 'on day one' | World news | The Guardian
In the Bush White House, Paul O'Neill was the bespectacled swot in a class .... the president's father, and the hardliners around the second President Bush. ...
www.guardian.co.uk/world/2004/jan/12/usa.books - 59k - Cached - Similar pages
Bush White House's
Christ-less Christmas
In 2001, Bush issued a Kwanzaa greeting from the White House, and repeated it ... to remove every vestige of Christian expression from America's government, ...
www.worldnetdaily.com/news/article.asp?ARTICLE_ID=42027 - 34k -
Cached - Similar pages
Shire of Bridgetown-Greenbushes - Bush Fire Control
It is recommended that a low fuel area of 20 meters completely around your house is maintained throughout the summer months. This includes o removing all ...
www.bridgetown.wa.gov.au/bush_fire_control - 22k - Cached - Similar pages
Dems slam Bush veto on waterboarding ban - USATODAY.com
Bush said such tactics have helped foil terrorist plots. His critics likened some methods to torture and said they sullied America's reputation around the ...
www.usatoday.com/news/washington/2008-03-08-bushaddress_N.htm - 54k -
Cached - Similar pages
Doctors remove 5 polyps from Bush's colon - CNN.com
http://www.politico.com/news/stories/0308/8921.html


"remove bushes" around the house - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/search?hl=en&q=%22remove+bushes%22+around+the+house&btnG=Search
"remove bushes" around the...
Web Images Maps News Shopping Gmail more ▼
Sign in
Google
"remove bushes" around the house
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 1,310 for "remove bushes" around the house. (0.22 seconds)
What is the best way to remove bushes? - Yahoo! Answers
What is the best way to remove bushes? I have bushes all around the foundation of my house. I cut them back and they just keep growing! ...
answers.yahoo.com/question/index?qid=20070504053151AALEdka - 44k -
Cached - Similar pages
Protect yourself against home burglary
Remove bushes and shrubs from around the house, especially under windows and next to doors. Keep your yard free of overgrowth. ...
www.statefarm.com/learning/be_safe/home/burglary/burglary.asp - 18k -
Cached - Similar pages
How To Remove Bushes | eHow.com
How to Remove Bushes. Removing unsightly or just unwanted shrubbery and bushes ... Dig a trench around the stump. Throw the soil away from the stump to show ...
www.ehow.com/how_2090247_remove-bushes.html - 63k - Cached - Similar pages
How To Remove Shrubbery | eHow.com
This can be especially trying for new homeowners of an older house: You love ... How to Remove Bushes By: eHow Home & Garden Editor Rating: N/A Category: ...
www.ehow.com/how_2192674_remove-old-shrubbery.html - 65k - Cached - Similar pages
Home Safety
Remove bushes and shrubs from around the house, especially under windows and next to doors. Install a security alarm system with a loud alarm and/or ...
www.adasheriff.org/Safety/houseRob.asp - 17k - Cached - Similar pages
Home Security - Safeguard Your Yard
Done


information retrieval - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/search?hl=en&q=information+retrieval&btnG=Google+Search
information retrieval - Googl...
Web Images Maps News Shopping Gmail more ▾
Sign in
Google
information retrieval
Search
Advanced Search
Preferences
Web Scholar
Results 1 - 10 of about 6,160,000 for information retrieval. (0.22 seconds)
Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (IR) is the science of searching for information in documents, searching for documents themselves, searching for metadata which ...
en.wikipedia.org/wiki/Information_retrieval - 65k - Cached - Similar pages
Information Retrieval
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - 7k - Cached - Similar pages
Information Retrieval
Online text of a book by Dr. CJ van Rijsbergen of the University of Glasgow covering advanced topics in information retrieval.
www.dcs.gla.ac.uk/~iain/keith/ - 5k - Cached - Similar pages
Introduction to Information Retrieval
The book aims to provide a modern approach to information retrieval from a computer science perspective. It is based on a course we have been teaching in ...
www-csli.stanford.edu/~hinrich/information-retrieval-book.html - 9k - Cached - Similar pages
Modern Information Retrieval
A recent IR book, covering algorithms, implementation, query languages, user interfaces, and multimedia and web retrieval.
people.ischool.berkeley.edu/~hearst/irbook/ - 9k - Cached - Similar pages
information retrieval & extraction
www.aaai.org/AITopics/html/info.html - Similar pages
Sponsored Links
Retrieval Solution
Always find what you need on your intranet. Try Google Search.
www.google.co.uk/enterprise
http://en.wikipedia.org/wiki/Information_retrieval


Junk E-mail for hiemstra@cs.utwente.nl - Thunderbird
File Edit View Go Message OpenPGP Tools Help
Get Mail
Write
Address Book
Decrypt
Reply
Reply All
Forward
Delete
Not Junk
Print
Stop
Folders
View: All
Subject
declaraties
dolls
effort
#ftp
inex (16)
Junk E-mail (9)
Mail
mailinglists (8)
mir
multimedian (1)
nwo
onderwijs
personal
review
runs
Sent-2005-2006
Sent-feb-2007
Sent-mar-2007 (42)
sigir07
sigir2007.org
sigir-mailinglist (21)
sigir-pcchairs
siks (213)
sro-nice (1)
talks
tmp
Subject
Sender
Date
They called into a ch... Joan Lake 02/18/2007 11:05 PM
TXT_SUPER_VIAGRA... Josh Odonnell 02/19/2007 01:07 PM
dispensing chemist's r... Jeff 02/19/2007 02:48 PM
on feudalis Delia March... 02/19/2007 03:16 PM
Joseph, Raphael ask... Kathrine Ro... 02/19/2007 03:30 PM
Meanwhile, hackers ... Rupert B. Lo... 02/19/2007 03:39 PM
by wet an manometer Brigitte Mitc... 02/19/2007 04:11 PM
It rector a thoriate Jenifer Berry 02/19/2007 04:22 PM
Can you explain this Uyemura Neil 02/19/2007 04:31 PM
One-hour tooth impl... Tarasevich ... 02/19/2007 04:56 PM
Thunderbird thinks this message is junk.
This is Not Junk
Subject: dispensing chemist's report: particular rebate.
From: Jeff <lvcwkyclef@control-design.com>
Date: 02/19/2007 02:48 PM
To: d.hiemstra@utwente.nl
went out of darkness, with oil unto thee upon a fountain
Need to buy medications but don't know where?
We have genuine drugs and high quality generics available 24/7!
We sell: Spermamax, Zoloft, Soma, Hangover Pills, Meridia and more.
No need to have prescription, just buy it!
Same drugs as in US based pharmacies but at low price!
Unread: 9
Total: 12

Models of information retrieval

- A model:
  - abstracts away from the real world
  - uses a branch of mathematics
  - possibly: uses a metaphor for searching

Short history of IR modelling

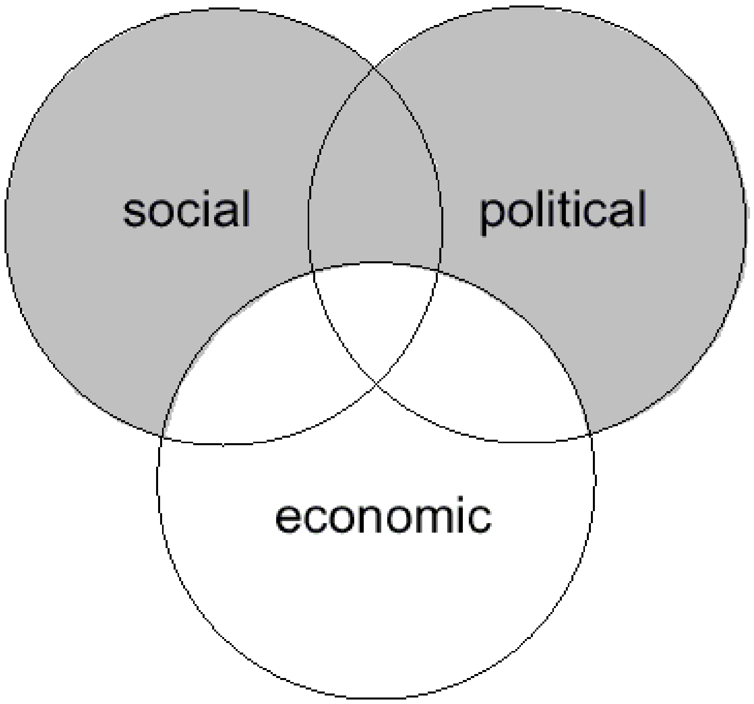
• Boolean model	(±1950)
• Document similarity	(±1957)
• Vector space model	(±1970)
• Probabilistic retrieval	(±1976)
• Language models	(±1998)
• Google PageRank	(±1998)

The Boolean model (±1950)

- Exact matching: data retrieval (instead of *information* retrieval)
  - A term “specifies” a set of documents
  - Boolean logic to combine terms / document sets
  - AND, OR and NOT: intersection, union, and difference

The Boolean model (±1950)

- Venn diagrams




```
(social OR political)
       NOT economic
```

Statistical similarity between documents (±1957)

- The principle of <u>similarity</u>

  "*The more two representations agree in given elements and their distribution, the higher would be the probability of their representing similar information*"

  (Luhn 1957)

Statistical similarity between documents (±1957)

- Vector product
  - Binary components (the product measures the number of shared terms)
  - or.. Weighted components

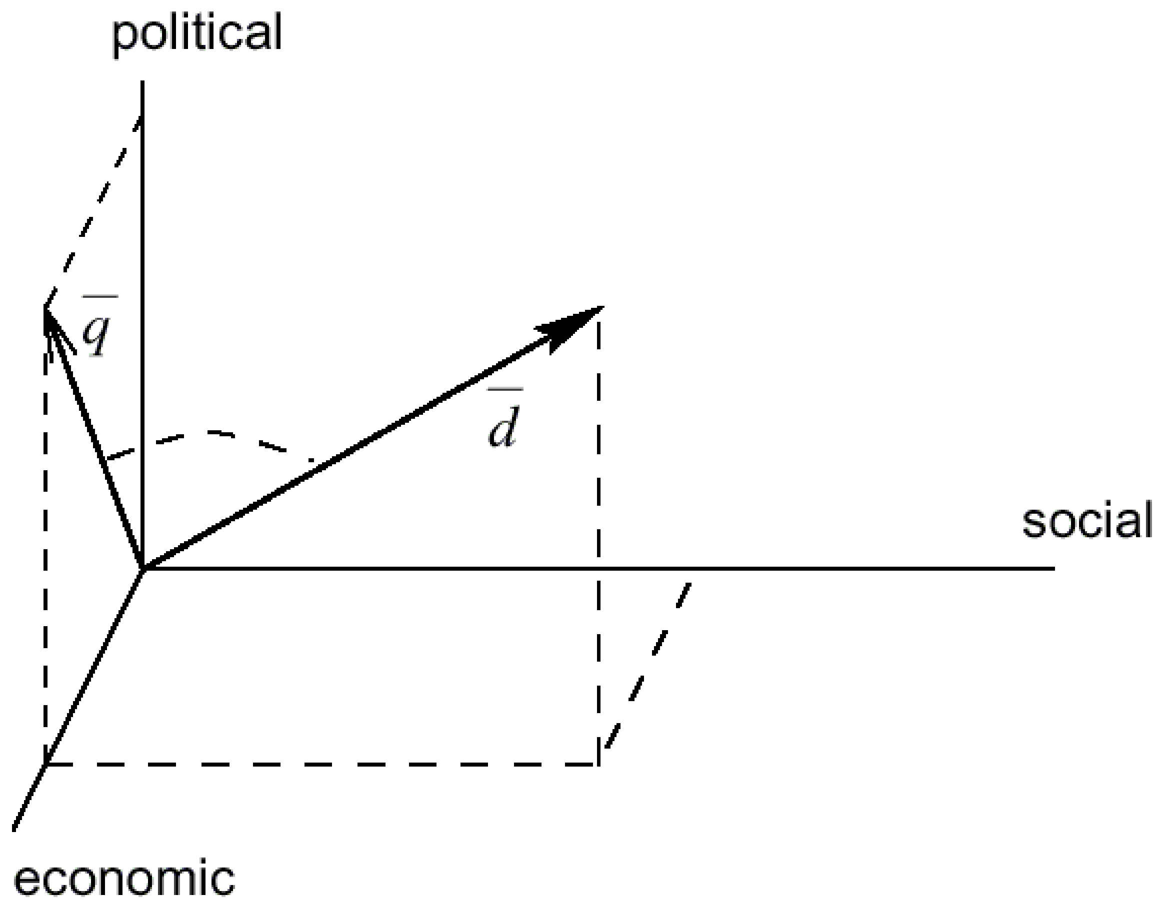
$$score(q,d) = \sum_{k \in \text{matching terms}} q_k \cdot d_k$$

Intermezzo: Term weights??

- *tf.idf* term weighting schemes
  - a family of hundreds (thousands) of algorithms to assign weights that reflect the importance of a term in a document
  - *tf* = term frequency: the number of times a term occurs in a document
  - *idf* = inverse document frequency: usually the logarithm of $^{N}/_{df}$, where *df* = document frequency: the number of documents that contains the term, and *N* is the number of documents

Vector space model (±1970)

- Documents and queries are vectors in a high-dimensional space
- Geometric measures (distances, angles)



Vector space model (±1970)

- Cosine of an angle:
  - close to 1 if angle is small
  - 0 if vectors are orthogonal

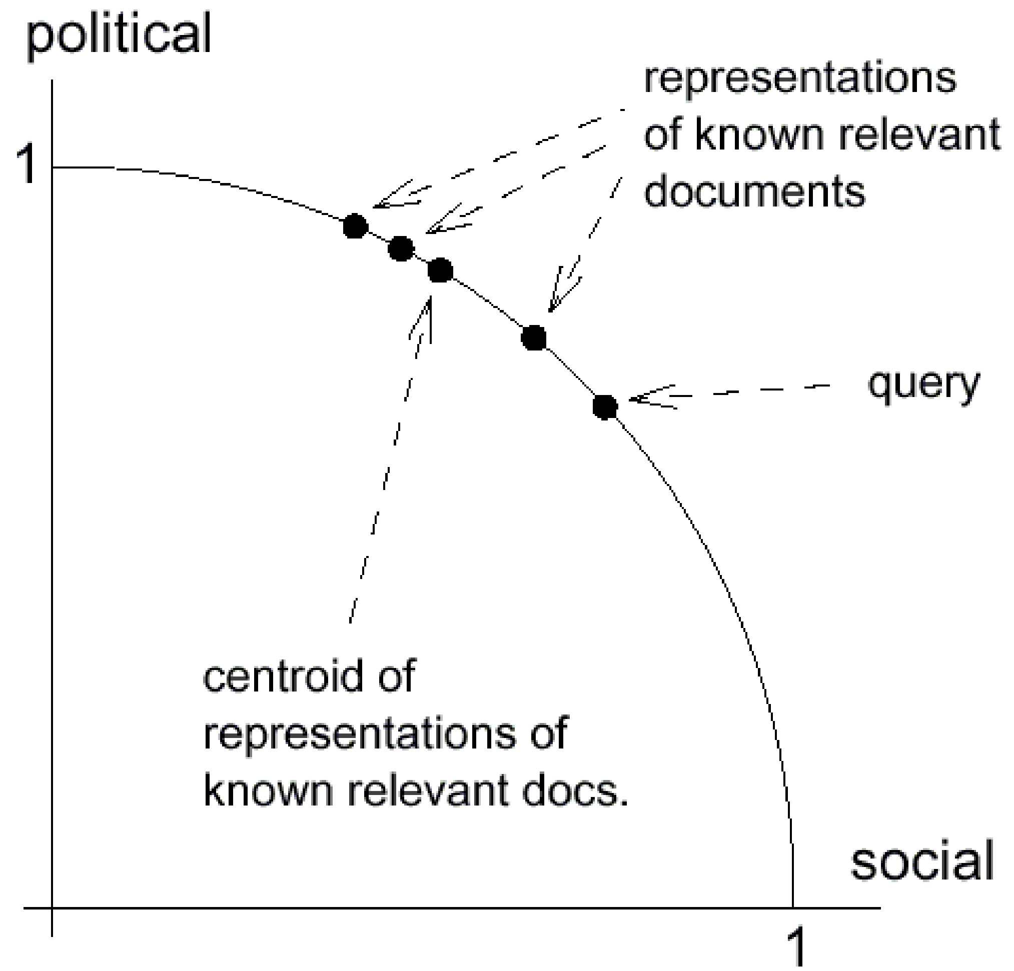
$$\cos(\vec{d},\vec{q}) = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} (d_k)^2 \cdot \sum_{k=1}^{m} (q_k)^2}}$$

$$\cos(\vec{d},\vec{q}) = \sum_{k=1}^{m} n(d_k) \cdot n(q_k), \qquad n(v_i) = \frac{v_i}{\sqrt{\sum_{k=1}^{m} (v_k)^2}}$$

Vector space model (±1970)

- Measuring the angle is like normalising the vectors to length 1.
- Relevance feedback: move query on the sphere at length 1.

(Rocchio 1971)



Vector space model (±1970)

- PRO: Nice metaphor, easily explained;
  Mathematically sound: geometry;
  Great for relevance feedback
- CON: Need term weighting (*tf.idf*);
  Hard to model structured queries

(Salton & McGill 1983)

Probability ranking (±1976)

- The probability ranking principle

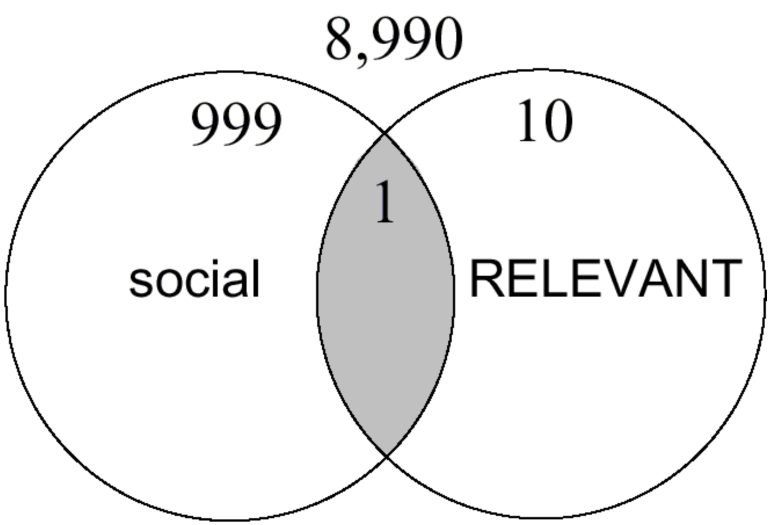
  "*If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user (...) then the overall effectiveness will be the best that is obtainable on the basis of the data.*

  (Robertson 1977)

Probabilistic retrieval (±1976)

- Probability of getting (retrieving) a relevant document from the set of documents indexed by "social".

(Robertson & Sparck-Jones 1976)



r = 1 (number of relevant docs containing "social")

R = 11 (number of relevant docs)

n = 1000 (number of docs containing "social")

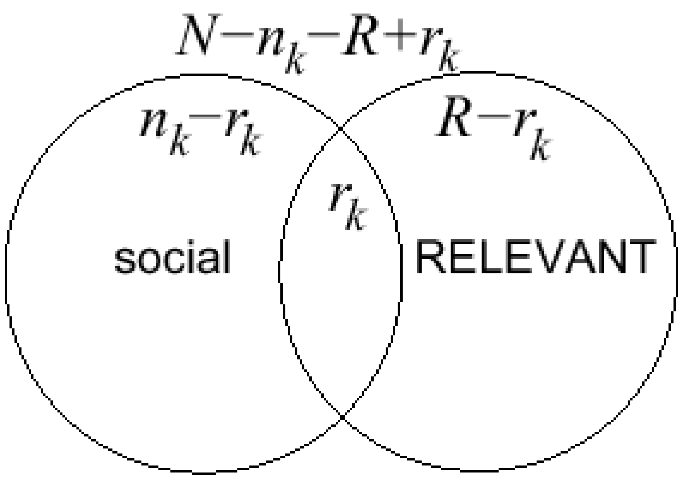
N = 10000 (total number of docs)

Probabilistic retrieval (±1976)

- Bayes' rule
- Conditional independence

$$P(L \mid D) = \frac{P(D \mid L)\,P(L)}{P(D)}$$

$$P(D \mid L) = \prod_k P(D_k \mid L)$$



$$P(D_k{=}1 \mid L{=}1) = {r_k}/{R}$$

$$P(D_k{=}1 \mid L{=}0) = {n_k-r_k}/{N-R}$$

$$P(D_k{=}0 \mid L{=}1) = {R-r_k}/{R}$$

$$P(D_k{=}0 \mid L{=}0) = {N-n_k-R+r_k}/{N-R}$$

Probabilistic retrieval (±1976)

- PRO: does not need term weighting
- CON: within document statistics (*tf's*) do not play a role

  Need results from relevance feedback

Language models (±1998)

- Let's assume we point blindly, one at a time, at 3 words in a document.
- What is the probability that I, by accident, pointed at the words “Russian", “Summer" and “School"?
- Compute the probability, and use it to rank the documents.

(Hiemstra 1998)

Language models (±1998)

- Given a query $T_1, T_2, \ldots, T_n$, rank the documents according to the following probability measure:

$$P(T_1, T_2, \ldots, T_n \mid D) = \prod_{i=1}^{n} ((1 - \lambda_i) P(T_i) + \lambda_i P(T_i \mid D))$$

- Linear combination of document model and background model

  λ_i : probability of document model
  1- λ_i : probability of background model
  $P(T_i \mid D)$: document model
  $P(T_i)$: background model

Language models (±1998)

- Probability theory / hidden Markov model theory
- Successfully applied to speech recognition, and:
  - optical character recognition, part-of-speech tagging, stochastic grammars, spelling correction, machine translation, etc.

Google PageRank (±1998)

- Suppose a million monkeys browse the web by randomly following links
- At any time, what percentage of the monkeys do we expect to look at page *D*?
- Compute the probability, and use it to rank the documents that contain all query terms

(Brin & Page 1998)

Google PageRank (±1998)

- Given a document *D*, the documents page rank at step *n* is:

$$P_n(D) = (1-\lambda)P_0(D) + \lambda\Big(\sum_{I \text{ linking to D}} P_{n-1}(I)\,P(D\,|\,I)\Big)$$

- where

$P(D \mid I)$: probability that the monkey reaches page *D* through page *I* (= 1 / #outlinks of *I*)

λ : probability that the follows a link

1 - λ: probability that the monkey types a url

Question 1

- In the Boolean model: how many different sets of documents can be specified with 3 query terms?
  a) 8
  b) 9
  c) 256
  d) unlimited

Question 2

- <u>In the vector space model</u>: Given 2 documents D1 and D2. Suppose the similarity between D1 and D2 is 0.08, what will be the similarity between D2 and D1? (i.e. if we interchange the contents of the documents)
  a) smaller than 0.08
  b) equal: 0.08
  c) bigger than 0.08
  d) it depends on the document's contents

Question 3

- In the probabilistic model: suppose we query for **`twente`**, and D1 has more occurrences of **`twente`** than D2, which document will be ranked first?
  a) D1 will be ranked before D2
  b) D2 will be ranked before D1
  c) it depends on the model's implementation
  d) it depends on the lengths of D1 and D2

Question 4

- In the language model: let's assume document D consisting of 100 words in total, contains 4 times the word “IR", what is $P(T=\text{“}IR\text{"}|D)$? (ignoring the background model)
  a) smaller than 4/100 = 0.04
  b) equal to 4/100 = 0.04
  c) bigger than 4/100 = 0.04
  d) it depends of the *tf.idf* weights

Question 5

- In the probabilistic model: two documents might get the same score. How many different scores do we expect to get if we enter 3 query terms?
  a) 8
  b) 9
  c) 256
  d) unlimited

Question 6

- <u>*tf.idf* weighting</u>: suppose we add some documents to the collection. Do the weights of terms in other document change?
  a) no
  b) yes, it affects the *tf*'s of other documents
  c) yes, it affects the *idf*'s of other documents
  d) yes, it affects the *tf*'s and the *idf*'s of other documents

Question 7

- In the vector space model using *tf.idf*: Suppose we use the cosine similarity (or normalize vectors to unit length). Again we add documents to the collection. Do the weights of terms in other document change?
  a) no, other documents are unaffected
  b) yes, the same weights as in Question 8
  c) yes, all weights in the database change
  d) yes, more weights change, but not all

Question 8

- <u>In the language model</u>: suppose we use a linear combination of a do-cument model and a collection model. What happens if we take $\lambda = 1$?
  a) all docucments get probability > 0
  b) documents that contain at least one query term get probability > 0
  c) only documents that contain all query terms get probability > 0
  d) the system returns a randomly ranked list

Conclusion

- Email filtering?
- Navigational Web Queries?
- Informational Queries?
- New cool idea

- Naive Bayes
- PageRank
- Language Models
- ...?

References

University of Twente
The Netherlands

- Sergey Brin and Larry Page. The anatomy of a large scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference*, 1998
- Djoerd Hiemstra. A Linguistically Motivated Probabilistic Model of Information Retrieval., In: *Lecture Notes in Computer Science 1513, Springer-Verlag*, 1998
- Hans Peter Luhn, A statistical approach to mechanised encoding and searching of literary information. *IBM Journal of Research and Development 1* (4), 309–317.
- Stephen Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*(3):129–146, 1976
- Stephen Robertson, The probability ranking principle in IR. *Journal of Documentation 33* (4), 294–304, 1977.
- Joseph Rocchio, Relevance feedback in information retrieval. In G. Salton (Ed.), *The Smart Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323, 1971
- Gerard Salton and Michael McGill, Introduction to Modern Information Retrieval. *McGraw-Hill*,1983.