UNIVERSITY OF TWENTE.

MODELS OF SEARCH

Foundations of Information Retrieval 2018

Djoerd Hiemstra

Dolf Trieschnigg

Theo Huibers

GOAL OF THIS LECTURE

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

COURSE MATERIAL (LAST WEEK)

- Introduction to Information Retrieval by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. Cambridge University Press. 2008. ISBN: 0521865719
    - Chapter 2, The term vocabulary & postings
    - Chapter 3, Dictionaries & tolerant retrieval
    - Chapter 6, Scoring & term weighting

COURSE MATERIAL (THIS WEEK)

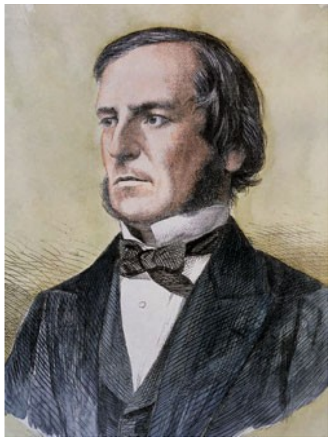
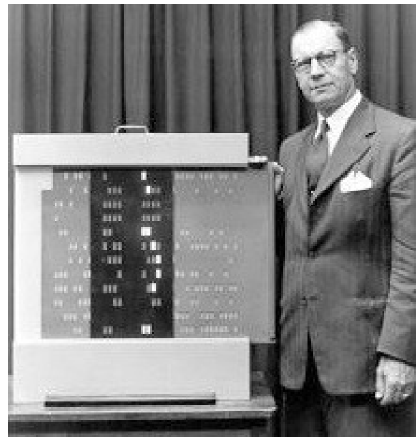
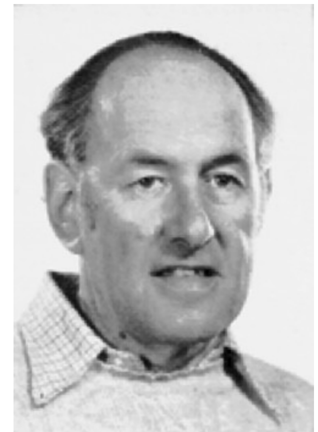
- Chapter 6, The Vector Space Model; Chapter 9, Relevance feedback & query expansion; Chapter 11, Probabilistic Information Retrieval; Chapter 12, Language Models; Chapter 21, Link Analysis

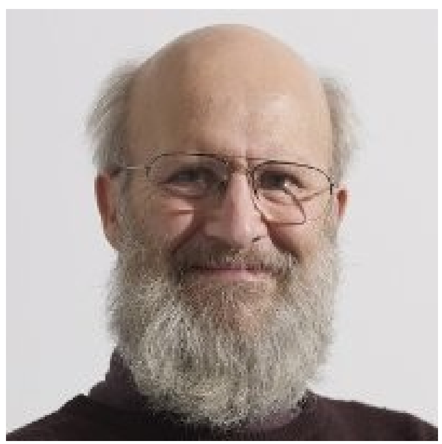
  http://informationretrieval.org

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

  http://www.cs.utwente.nl/~hiemstra/papers/IRModelsTutorial-draft.pdf

NOTABLE PEOPLE



George Boole, Hans Peter Luhn, Gerard Salton, Karen Sparck-Jones, Stephen Robertson, Frederick Jelinek, Larry Page

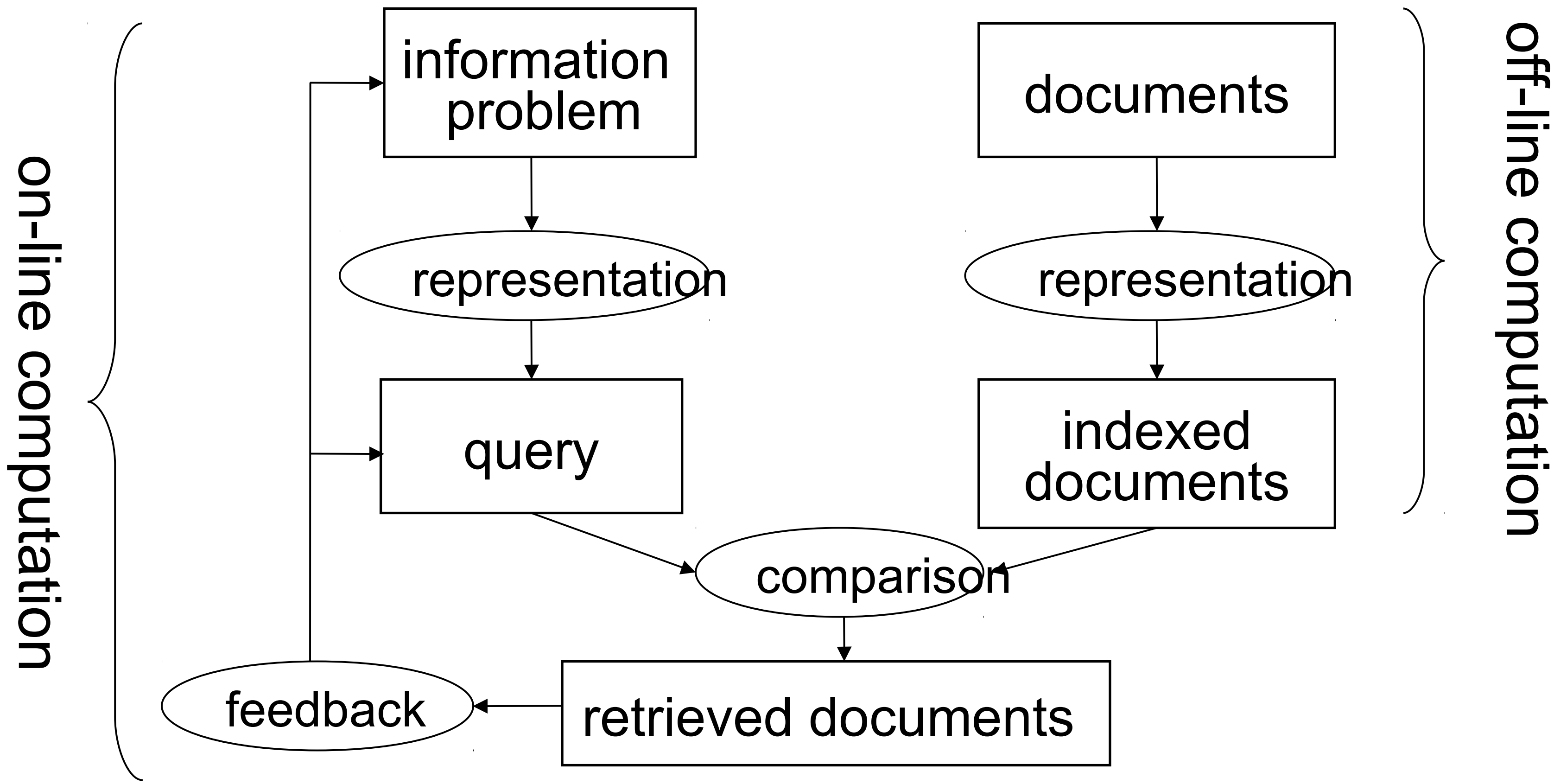
OVERVIEW

- PART 1: Looking back
- PART 2: IR models
  - Basic technology
  - An overview of formal models
- PART 3: The Quiz


Is this going to be another "According To Google" answer?
© Philipp Lenssen
blog.outer-court.com

PART-1: LOOKING BACK

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, hyperlinks)
  - 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:
  - tokenization
  - statistics (counting words)
  - stop list
  - morphological stemming
  - compound splitting
  - partial parsing: noun phrase extraction
  - other: use of thesaurus/synonyms, 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
  - What about “bruidsluier”?
- phrases
  - separate words not good predictors: New York

BEING AN IR SYSTEM

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

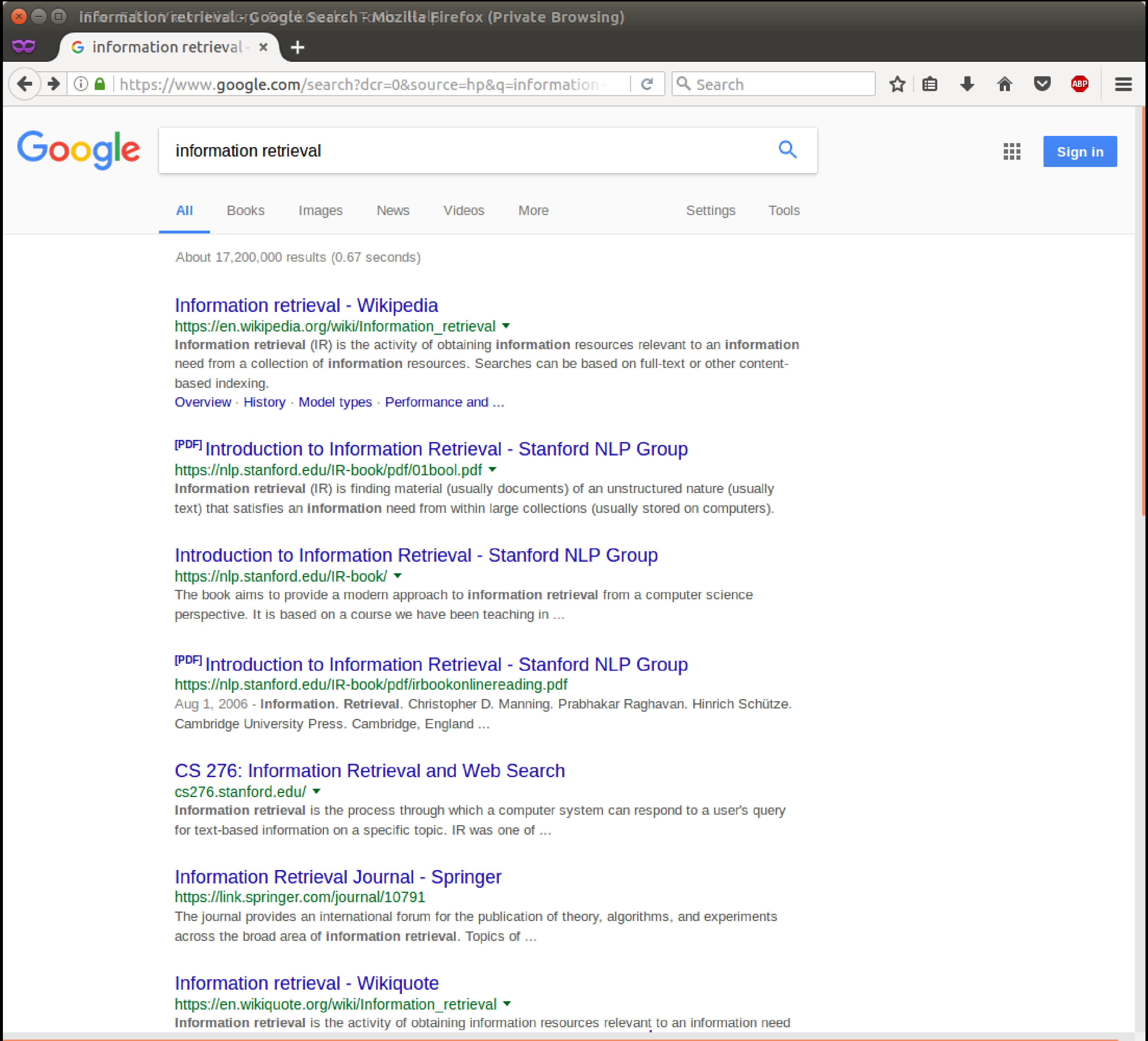
Massachusetts dumps Microsoft Office

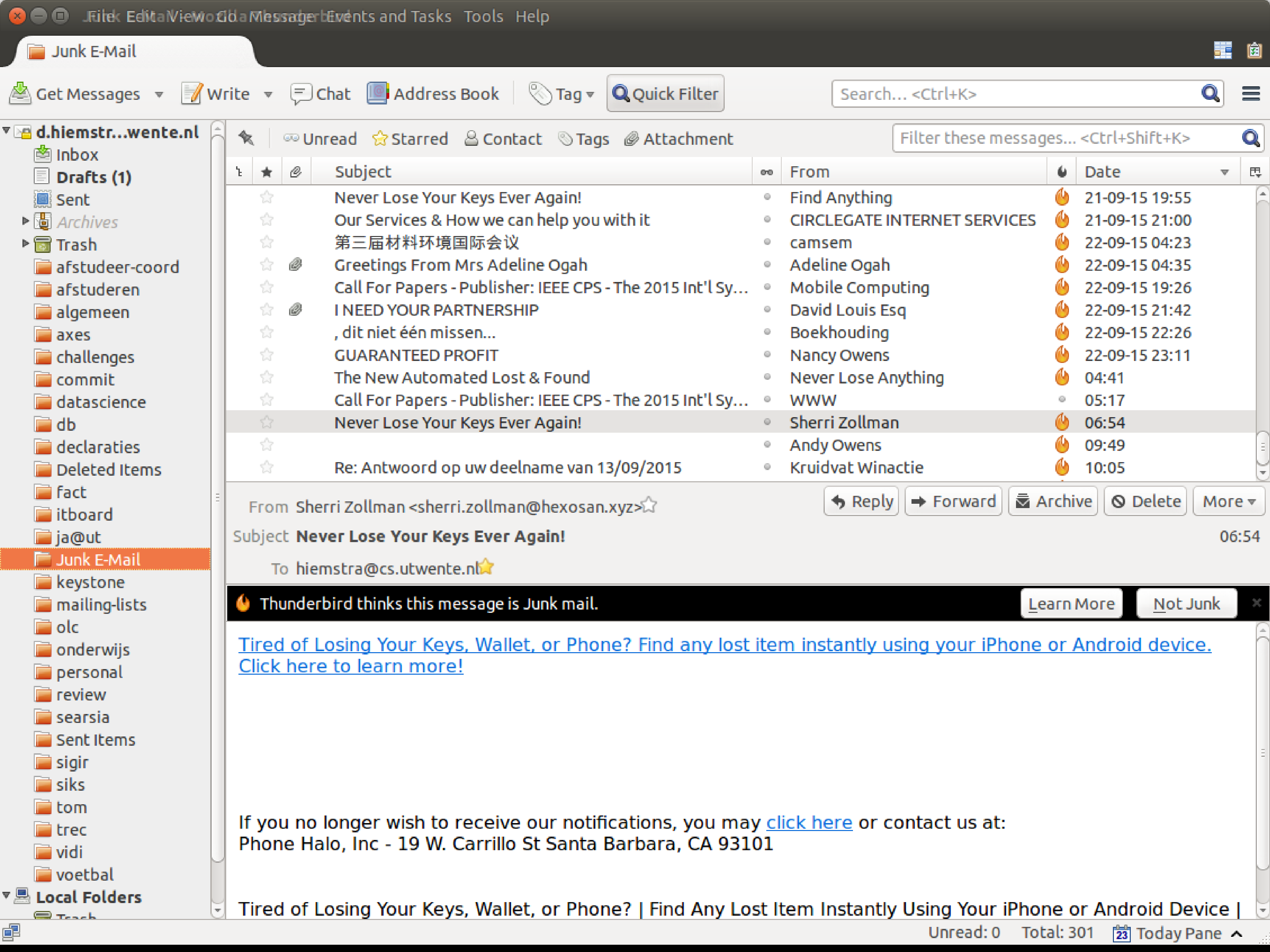
The people who brought you the Boston tea party, have joined in another revolution against good King Billy's Office software. The state government has decided that all electronic documents saved and created by state employees have to use open formats.

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 SYSTEM

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.


information retrieval - Google Search - Mozilla Firefox (Private Browsing)
information retrieval -
https://www.google.com/search?dcr=0&source=hp&q=information+
Search
Google
information retrieval
Sign in
All
Books
Images
News
Videos
More
Settings
Tools
About 17,200,000 results (0.67 seconds)
Information retrieval - Wikipedia
https://en.wikipedia.org/wiki/Information_retrieval
Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.
Overview · History · Model types · Performance and ...
[PDF] Introduction to Information Retrieval - Stanford NLP Group
https://nlp.stanford.edu/IR-book/pdf/01bool.pdf
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
Introduction to Information Retrieval - Stanford NLP Group
https://nlp.stanford.edu/IR-book/
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 ...
[PDF] Introduction to Information Retrieval - Stanford NLP Group
https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
Aug 1, 2006 - Information. Retrieval. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. Cambridge University Press. Cambridge, England ...
CS 276: Information Retrieval and Web Search
cs276.stanford.edu/
Information retrieval is the process through which a computer system can respond to a user's query for text-based information on a specific topic. IR was one of ...
Information Retrieval Journal - Springer
https://link.springer.com/journal/10791
The journal provides an international forum for the publication of theory, algorithms, and experiments across the broad area of information retrieval. Topics of ...
Information retrieval - Wikiquote
https://en.wikiquote.org/wiki/Information_retrieval
Information retrieval is the activity of obtaining information resources relevant to an information need


File Edit View Go Message Events and Tasks Tools Help
Junk E-Mail
Get Messages
Write
Chat
Address Book
Tag
Quick Filter
Search… <Ctrl+K>
d.hiemstr...wente.nl
Inbox
Drafts (1)
Sent
Archives
Trash
afstudeer-coord
afstuderen
algemeen
axes
challenges
commit
datascience
db
declaraties
Deleted Items
fact
itboard
ja@ut
Junk E-Mail
keystone
mailing-lists
olc
onderwijs
personal
review
searsia
Sent Items
sigir
siks
tom
trec
vidi
voetbal
Local Folders
Unread
Starred
Contact
Tags
Attachment
Filter these messages… <Ctrl+Shift+K>
Subject
From
Date
Never Lose Your Keys Ever Again! | Find Anything | 21-09-15 19:55
Our Services & How we can help you with it | CIRCLEGATE INTERNET SERVICES | 21-09-15 21:00
第三届材料环境国际会议 | camsem | 22-09-15 04:23
Greetings From Mrs Adeline Ogah | Adeline Ogah | 22-09-15 04:35
Call For Papers - Publisher: IEEE CPS - The 2015 Int'l Sy… | Mobile Computing | 22-09-15 19:26
I NEED YOUR PARTNERSHIP | David Louis Esq | 22-09-15 21:42
, dit niet één missen… | Boekhouding | 22-09-15 22:26
GUARANTEED PROFIT | Nancy Owens | 22-09-15 23:11
The New Automated Lost & Found | Never Lose Anything | 04:41
Call For Papers - Publisher: IEEE CPS - The 2015 Int'l Sy… | WWW | 05:17
Never Lose Your Keys Ever Again! | Sherri Zollman | 06:54
Andy Owens | 09:49
Re: Antwoord op uw deelname van 13/09/2015 | Kruidvat Winactie | 10:05
From Sherri Zollman <sherri.zollman@hexosan.xyz>
Reply
Forward
Archive
Delete
More
Subject Never Lose Your Keys Ever Again!
06:54
To hiemstra@cs.utwente.nl
Thunderbird thinks this message is Junk mail.
Learn More
Not Junk
Tired of Losing Your Keys, Wallet, or Phone? Find any lost item instantly using your iPhone or Android device.
Click here to learn more!
If you no longer wish to receive our notifications, you may click here or contact us at:
Phone Halo, Inc - 19 W. Carrillo St Santa Barbara, CA 93101
Tired of Losing Your Keys, Wallet, or Phone? | Find Any Lost Item Instantly Using Your iPhone or Android Device |
Unread: 0
Total: 301
Today Pane

PART-2: INFORMATION RETRIEVAL MODELS

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 MODELING

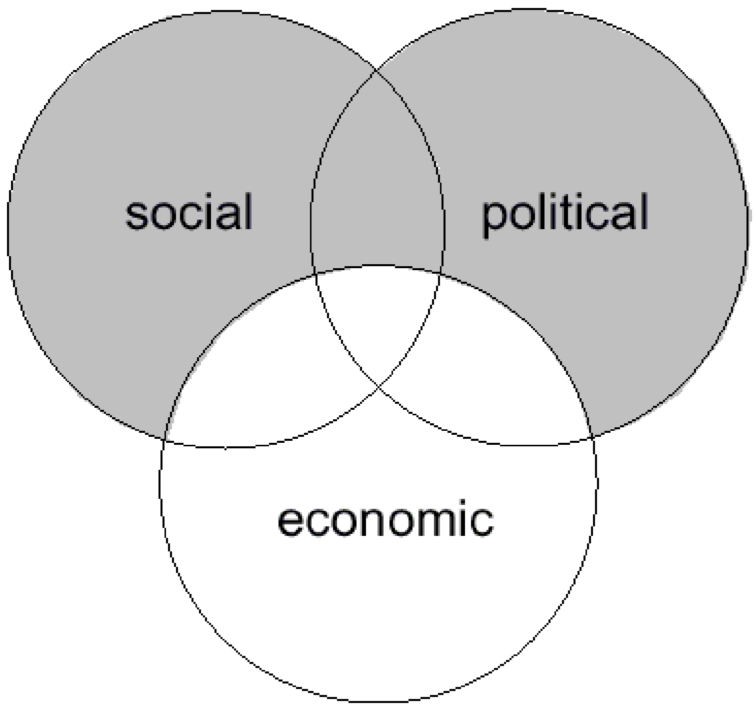
- Boolean model (±1950)
- Document similarity (±1957)
- Vector space model (±1970)
- Probabilistic retrieval (±1976)
- Language models (±1998)
- Google PageRank (±1998)

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 similarity

  "*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
  - If the vector has binary components, then the product measures the number of shared terms
  - Vector components might be "weights"

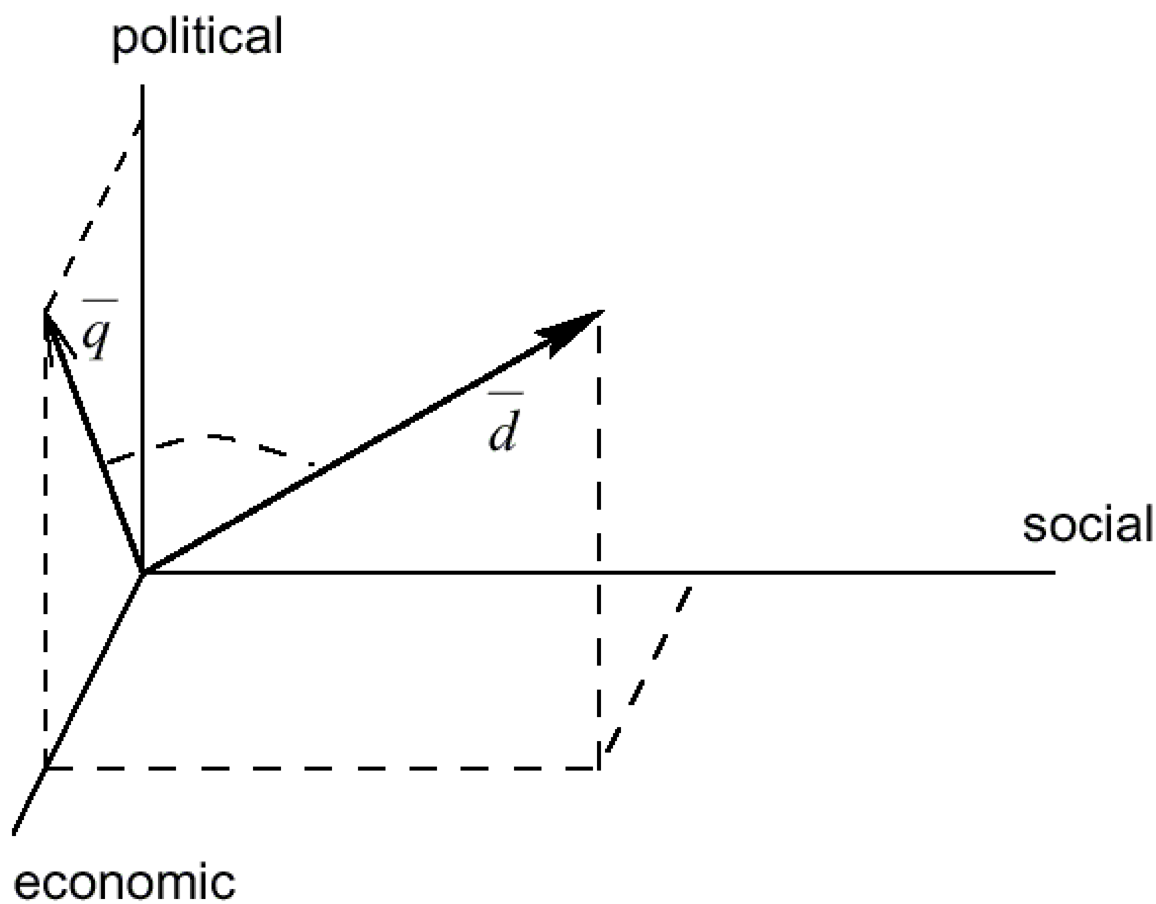
$$score(\vec{q}, \vec{d}) = \sum_{k \in \text{matching terms}} q_k \cdot d_k$$

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

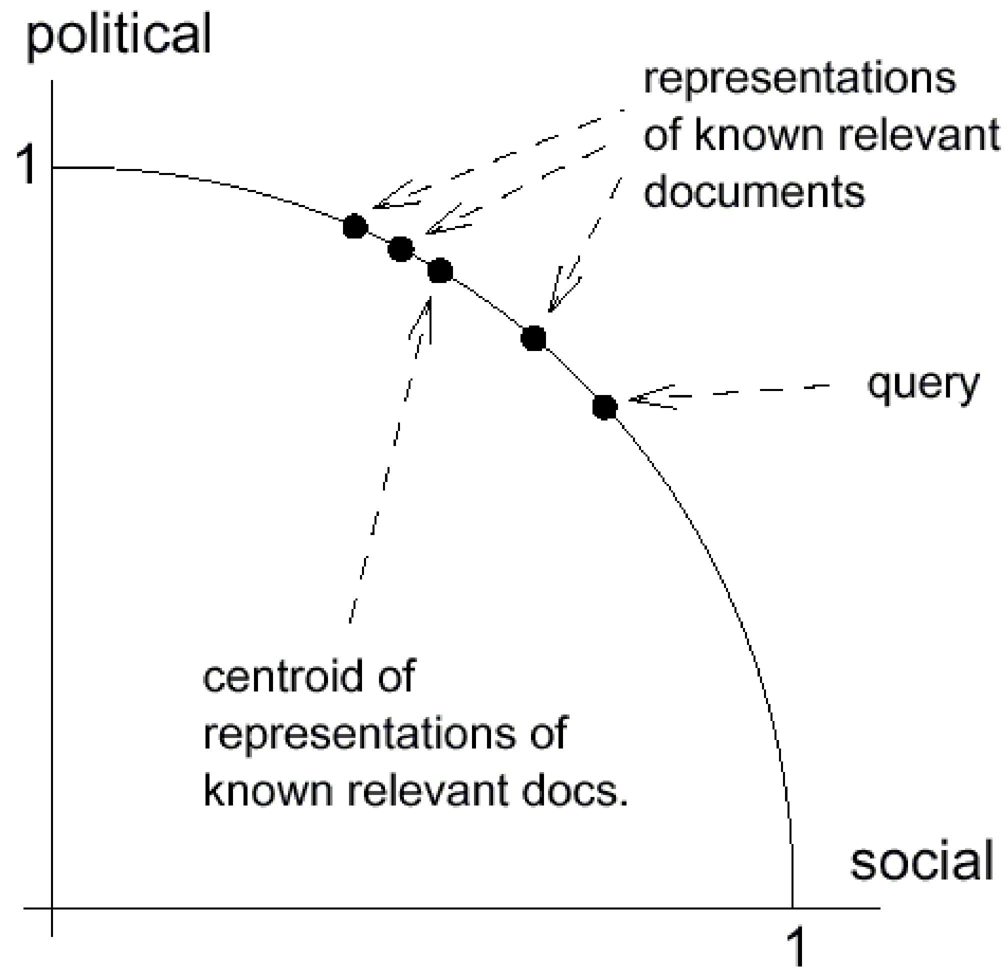
$$\cos(\vec{d},\vec{q}) = \frac{\sum_{k=1}^{m} d_k \cdot q_k}{\sqrt{\sum_{k=1}^{m} (d_k)^2 \cdot \sum_{k=1}^{m} (q_k)^2}}$$

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

VECTOR SPACE MODEL (±1970)

- Measuring the angle is like normalising 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

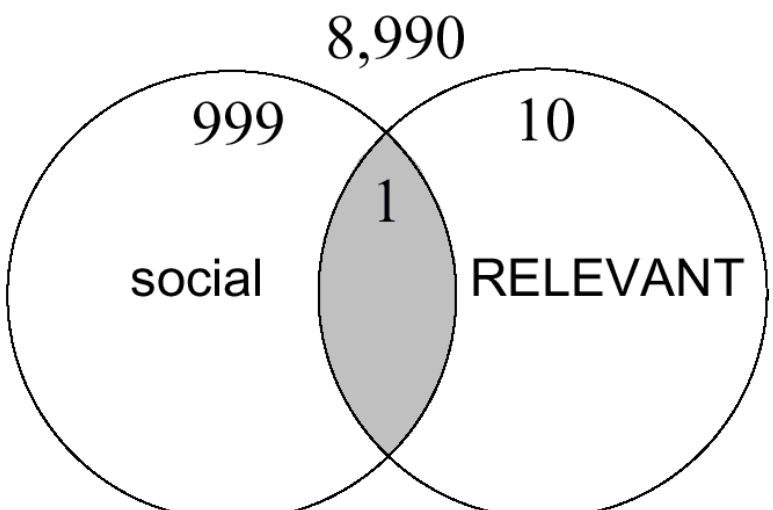
  "*If a reference retrieval system's response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user (...) then the overall effectiveness will be the best that is obtainable on the basis of the data.*

  (Robertson 1977)

PROBABILITY RANKING (±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")

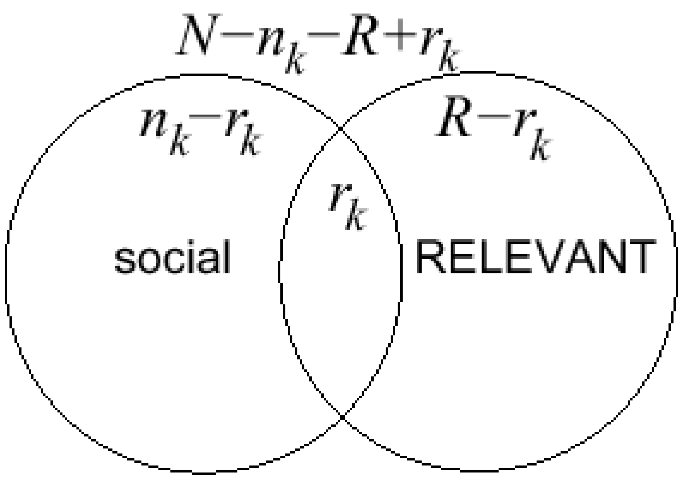
N = 10000 (total number of docs)

PROBABILITY RANKING (±1976)

- Bayes' rule
- Conditional independence

$$P(L|D)=\frac{P(D|L)P(L)}{P(D)}$$

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



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

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

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

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

PROBABILITY RANKING (±1976)

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

  Need results from relevance feedback

(Trivia: also known as BM1)

OKAPI BM25 (±1994)

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D) + k_1 \cdot \left(1 - b + b\cdot \frac{|D|}{\text{avgdl}}\right)},$$

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

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 “Master", “Computer" and “Science"?
- Compute the probability, and use it to rank the documents.

LANGUAGE MODELS (±1998)

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

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

- Linear combination of document model and background model

λ :	probability of document model
$1\text{-}\lambda$:	probability of background model
$P(T_i \mid D)$:	document model
$P(T_i)$:	background model

Jelinek-Mercer Smoothing?

Frederick Jelinek and Robert Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In: Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam.

LANGUAGE MODELS (±1998)

- $$P\left(D \mid T_1,\ldots,T_n\right)=\frac{P\left(T_1,\ldots,T_n \mid D\right)P\left(D\right)}{P\left(T_1,\ldots,T_n\right)}$$
- 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.

LANGUAGE MODELS (±1998)

- A whole family of models
  - Document priors
  - Relevance models (pseudo feedback)
  - Translation models (cross-language)
  - Aspect models (latent semantic indexing)

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

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

$1\text{-}\lambda$: probability that the monkey types a url

Managing Big Data (201200044)

- M. CS (k2)

  The course will closely follow developments to manage big data on large clusters of commodity machines, initiated by Google, and followed by many other web companies such as Yahoo, Amazon, Facebook, Spotify, Twitter, etc. Big data gives rise to a redesign of many core computer science concepts: The course discusses file systems (Google FS), programming paradigms (MapReduce), programming languages and query languages (Spark and Pig Latin), 'noSQL' database paradigms (for instance Google's BigTable) for managing big data, and solutions for managing streaming data (for instance Twitter's Storm).

UNIVERSITY OF TWENTE.

PART-3:
THE FORMAL MODEL
QUIZ

QUESTION 1

<u>In the Boolean model</u>: 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

<u>In the probabilistic model</u>: 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

<u>In the language model</u>: 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 ex-pect 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

In a language model: suppose we use a linear combination of a document 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

- There is *no* standard theory for building information retrieval systems
  - unlike e.g. databases: relational model
  - so, no standard query language
- Many issues hardly addressed by models
  - ranking with structured queries
  - ranking with structured documents
  - non-content information (e.g. Google PageRank)
  - combining media: e.g. textual and feature-based queries