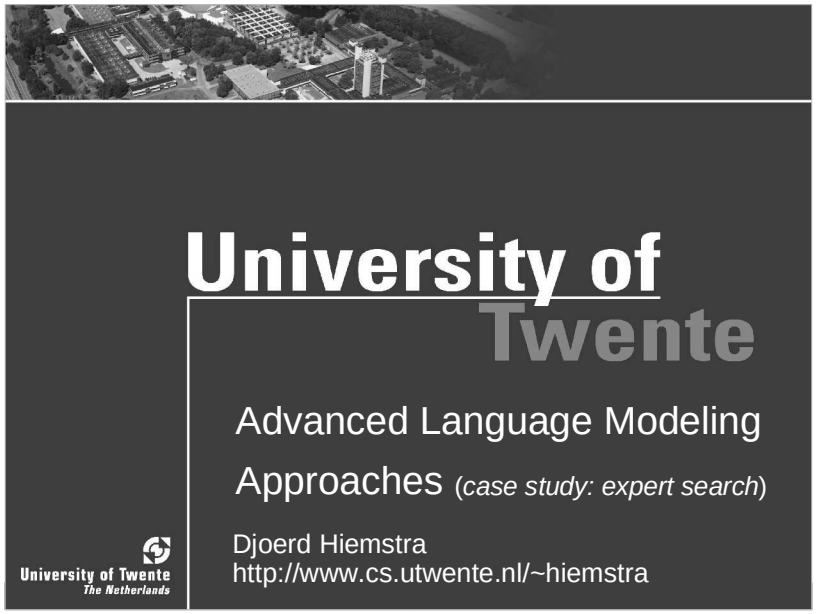
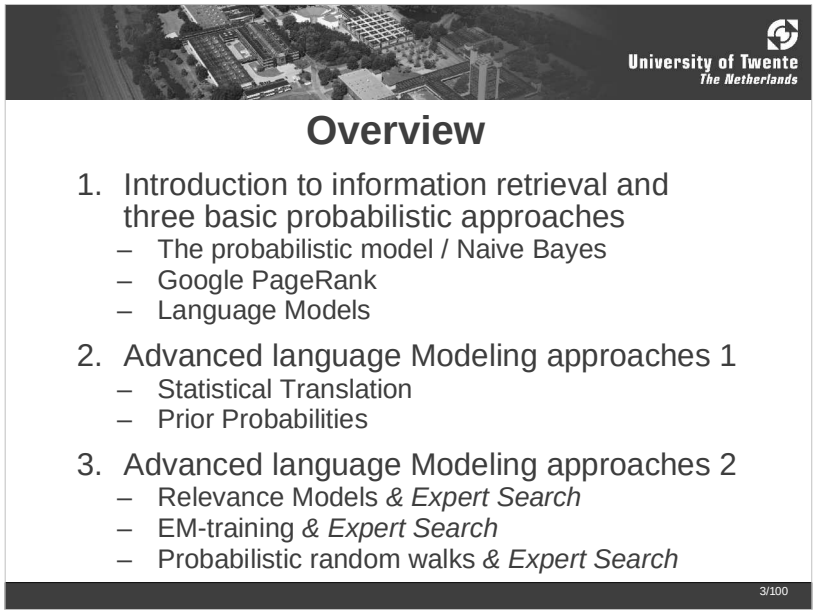
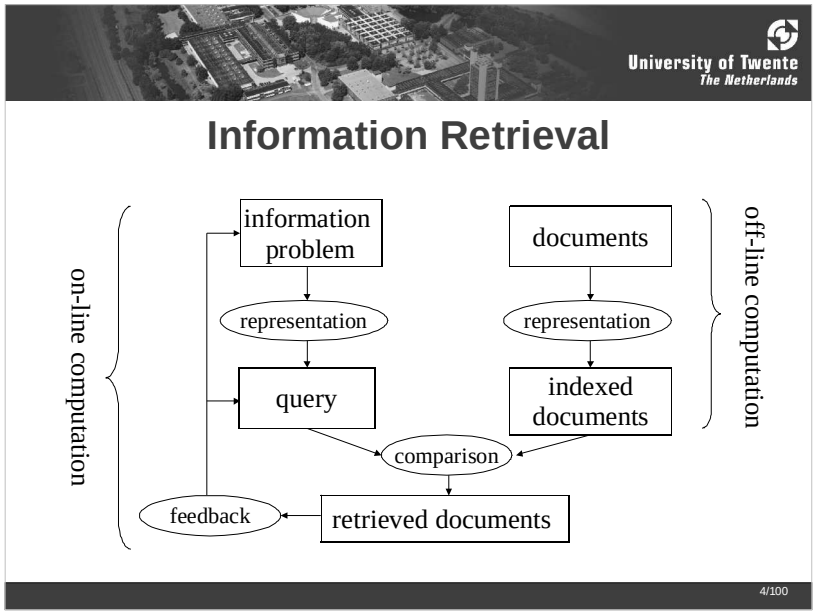
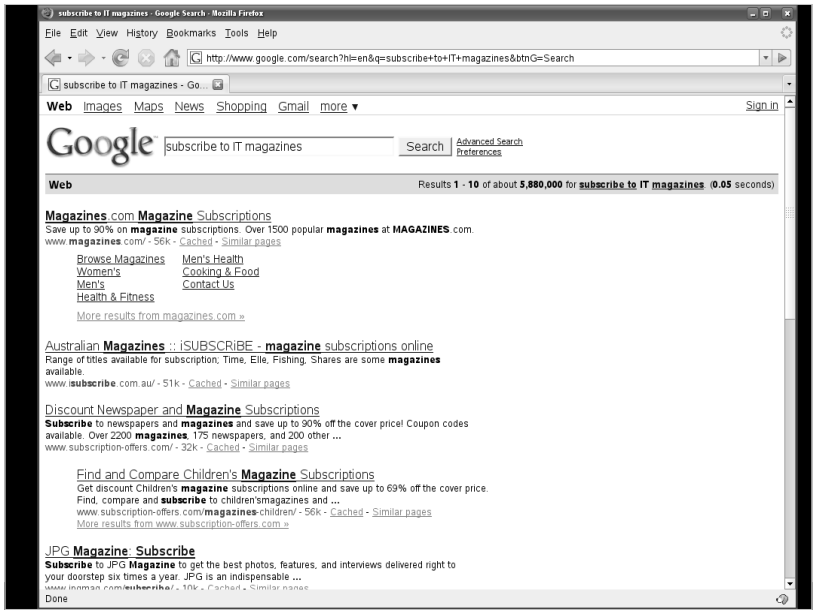
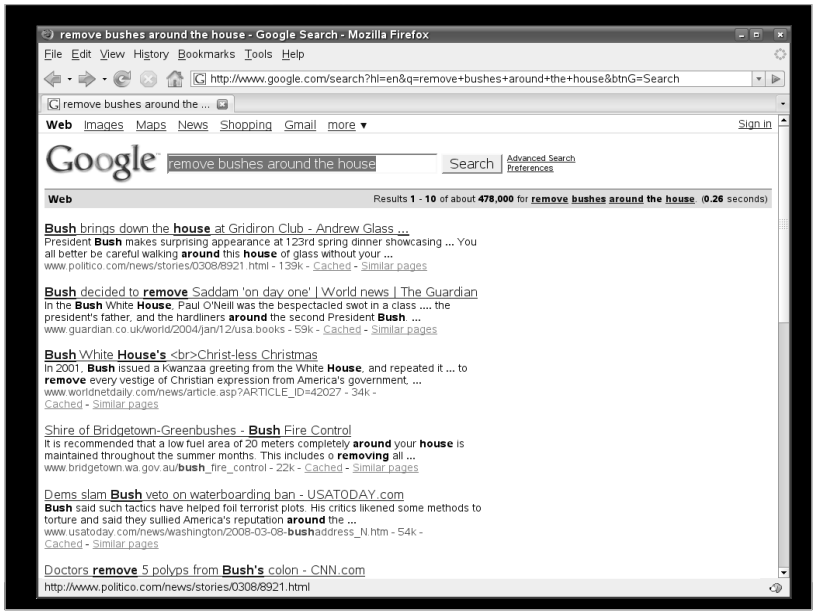
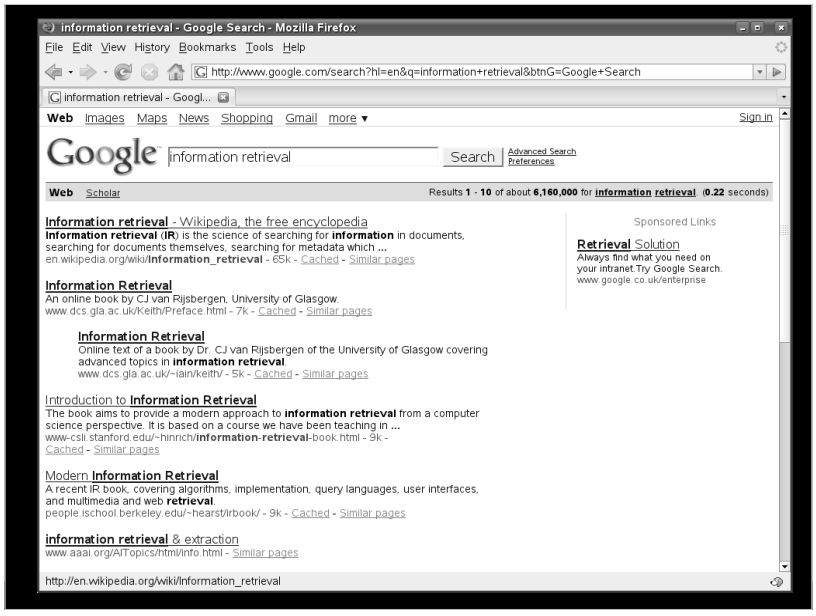

University of
Twente
Advanced Language Modeling
Approaches (case study: expert search)
Djoerd Hiemstra
http://www.cs.utwente.nl/~hiemstra
University of Twente
The Netherlands

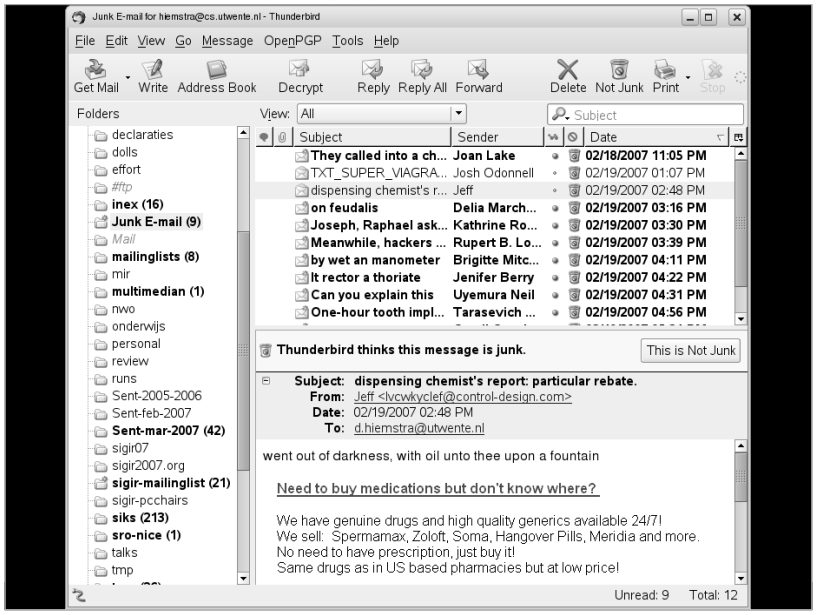

University of Twente
The Netherlands
Overview
1. Introduction to information retrieval and three basic probabilistic approaches
– The probabilistic model / Naive Bayes
– Google PageRank
– Language Models
2. Advanced language Modeling approaches 1
– Statistical Translation
– Prior Probabilities
3. Advanced language Modeling approaches 2
– Relevance Models & Expert Search
– EM-training & Expert Search
– Probabilistic random walks & Expert Search
3/100

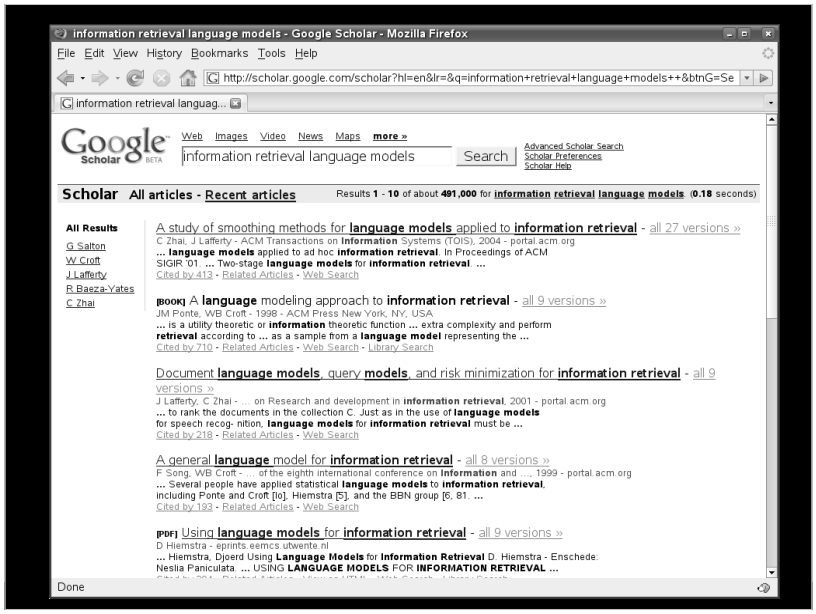

University of Twente
The Netherlands
Information Retrieval
information problem
representation
query
documents
representation
indexed documents
comparison
retrieved documents
feedback
on-line computation
off-line computation
4/100


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
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
Women's
Men's
Health & Fitness
Men's Health
Cooking & Food
Contact Us
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.isubscribe.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 ...
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
Google
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 <br>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


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
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
View: All
Subject
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


information retrieval language models - Google Scholar - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://scholar.google.com/scholar?hl=en&lr=&q=information+retrieval+language+models++&btnG=Se
information retrieval languag...
Google Scholar BETA
Web Images Video News Maps more »
information retrieval language models
Search
Advanced Scholar Search
Scholar Preferences
Scholar Help
Scholar All articles - Recent articles
Results 1 - 10 of about 491,000 for information retrieval language models (0.18 seconds)
All Results
G Salton
W Croft
J Lafferty
R Baeza-Yates
C Zhai
A study of smoothing methods for language models applied to information retrieval - all 27 versions »
C Zhai, J Lafferty - ACM Transactions on Information Systems (TOIS), 2004 - portal.acm.org
... language models applied to ad hoc information retrieval. In Proceedings of ACM
SIGIR '01. ... Two-stage language models for information retrieval. ...
Cited by 413 - Related Articles - Web Search
[BOOK] A language modeling approach to information retrieval - all 9 versions »
JM Ponte, WB Croft - 1998 - ACM Press New York, NY, USA
... is a utility theoretic or information theoretic function ... extra complexity and perform
retrieval according to ... as a sample from a language model representing the ...
Cited by 710 - Related Articles - Web Search - Library Search
Document language models, query models, and risk minimization for information retrieval - all 9 versions »
J Lafferty, C Zhai - ... on Research and development in information retrieval, 2001 - portal.acm.org
... to rank the documents in the collection C. Just as in the use of language models
for speech recog- nition, language models for information retrieval must be ...
Cited by 218 - Related Articles - Web Search
A general language model for information retrieval - all 8 versions »
F Song, WB Croft - ... of the eighth international conference on Information and ..., 1999 - portal.acm.org
... Several people have applied statistical language models to information retrieval,
including Ponte and Croft [6], Hiemstra [5], and the BBN group [6, 8]. ...
Cited by 193 - Related Articles - Web Search
[PDF] Using language models for information retrieval - all 9 versions »
D Hiemstra - eprints.eemcs.utwente.nl
... Hiemstra, Djoerd Using Language Models for Information Retrieval D. Hiemstra - Enschede:
Neslia Paniculata. ... USING LANGUAGE MODELS FOR INFORMATION RETRIEVAL ...
Done


# PART 1
# Introduction to probabilistic information retrieval

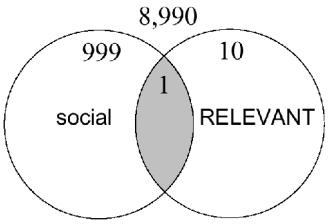



## IR Models: probabilistic models

- Rank documents by the probability that, for instance:
  - A random document from the documents that contain the query is relevant (known as “the probabilistic model” or “naïve Bayes”)
  - A random surfer visits the page (known as “Google PageRank”)
  - Random words from the document form the query (known as “language models”)


## Probabilistic model (Robertson & Sparck-Jones 1976)

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




$r = 1$ (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)

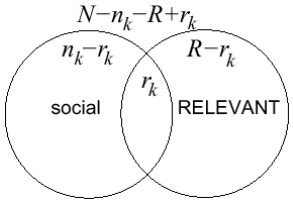
14/100


## Probabilistic retrieval

- Bayes' rule
- Conditional independence

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

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




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

15/100


## Google PageRank (Brin & Page 1998)

- Suppose a million monkeys browse the www 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

- 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 $I$)
  - $\lambda$: probability that the follows a link
  - $1-\lambda$: probability that the monkey types a url


## Language models?

- A language model assigns a probability to a piece of language (i.e. a sequence of tokens)

  $P$(how are you today)

  >

  $P$(cow barks moo souflé)

  >

  $P$(asj mokplah qnbgol yokii)




## Language models (Hiemstra 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 “ECIR", “models" and “tutorial"?
- Compute the probability, and use it to rank the documents.


## Language models

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



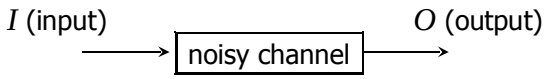

## Half way conclusion

- Email filtering?
- Navigational Web Queries?
- Informational Queries?
- Expert Search?
- Naive Bayes
- PageRank
- Language Models
- ...


# PART 2
# Advanced statistical language models



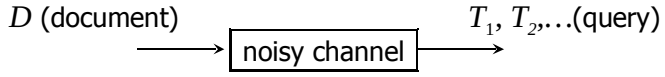

## Noisy channel paradigm (Shannon 1948)




- hypothesise all possible input texts $I$ and take the one with the highest probability, symbolically:

$$\hat{I} = \arg\max_{I} P(I|O)$$
$$= \arg\max_{I} P(I) \cdot P(O|I)$$


## Noisy channel paradigm (Shannon 1948)




- hypothesise all possible documents $D$ and take the one with the highest probability, symbolically:

$$\hat{D} = \arg\max_{D} P(D|T_1, T_2, \cdots)$$
$$= \arg\max_{D} P(D) \cdot P(T_1, T_2, \cdots|D)$$




## Noisy channel paradigm

- Did you get the picture? Formulate the following systems as a noisy channel:
  - Automatic Speech Recognition
  - Optical Character Recognition
  - Parsing of Natural Language
  - Machine Translation
  - Part-of-speech tagging


## Statistical language models

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

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

$\lambda_i$ : probability that the term on position *i* is important

$1-\lambda_i$ : probability that the term is unimportant

$P(T_i | D)$ : probability of an important term

$P(T_i)$ : probability of an unimportant term




## Statistical language models

- Definition of probability measures:

$$P(T_i = t_i | D = d) = \frac{tf(t_i, d)}{\sum_t tf(t, d)} \quad \text{(important term)}$$

$$P(T_i = t_i) = \frac{df(t_i)}{\sum_t df(t)} \quad \text{(unimportant term)}$$

$\lambda_i = 0.5$


## Exercise: an expert search test collection

1. Define your personal three-word language model:
   Choose three words (and for each word a probability)
2. Write two joint papers, each with two or more co-authors of your choice for the Int. Conference on Short Papers (ICSP)
   - Papers must not exceed two words per author
   - Use only words from your personal language model
   - ICSP does not do blind reviewing, so clearly put the names of the authors on the paper
   - Deadline: after the coffee-break.
3. Question: Can the PC find out who are experts on *x*?




## Exercise 2: simple LM scoring

- Calculate the language modeling scores for the query *y* on your document(s)
  - What needs to be decided before we are able to do this?
  - 5 minutes!

University of Twente
The Netherlands

## Statistical language models

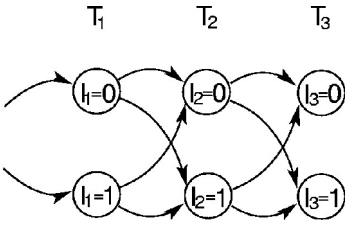
- How to estimate value of $\lambda_i$ ?
  - For ad-hoc retrieval (i.e. no previously retrieved documents to guide the search)
    $\lambda_i = constant$ (i.e. each term equally important)
  - Note that for extreme values:
    $\lambda_i = 0$ : term does not influence ranking
    $\lambda_i = 1$ : term is mandatory in retrieved docs.
    $\lim \lambda_i \to 1$ : docs containing $n$ query terms are ranked above docs containing $n - 1$ terms (Hiemstra 2004)

31/100

University of Twente
The Netherlands

## Statistical language models

- Presentation as hidden Markov model
  - finite state machine: probabilities governing transitions
  - sequence of state transitions cannot be determined from sequence of output symbols (i.e. are hidden)




32/100


## Statistical language models

- Implementation

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

$$\vdots$$

$$P(T_1, T_2, \cdots, T_n|D) \propto \sum_{i=1}^{n} \log(1 + \frac{\lambda_i P(T_i|D)}{(1-\lambda_i)P(T_i)})$$




## Statistical language models

- Implementation as vector product:

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

$$q_k = tf(k,q)$$

$$d_k = \log(1 + \frac{tf(k,d) \sum_t df(t)}{df(k) \sum_t tf(t,d)} \cdot \frac{\lambda_k}{1-\lambda_k})$$

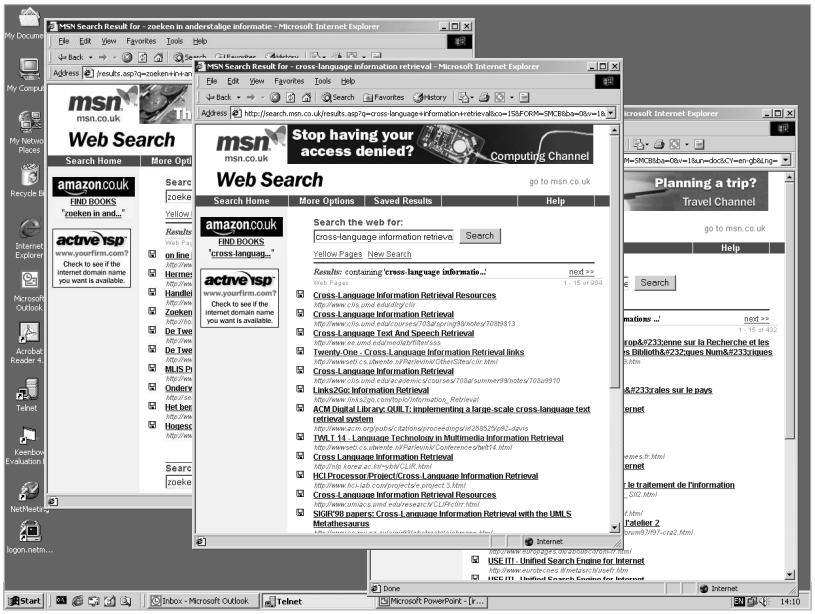

## Cross-language IR

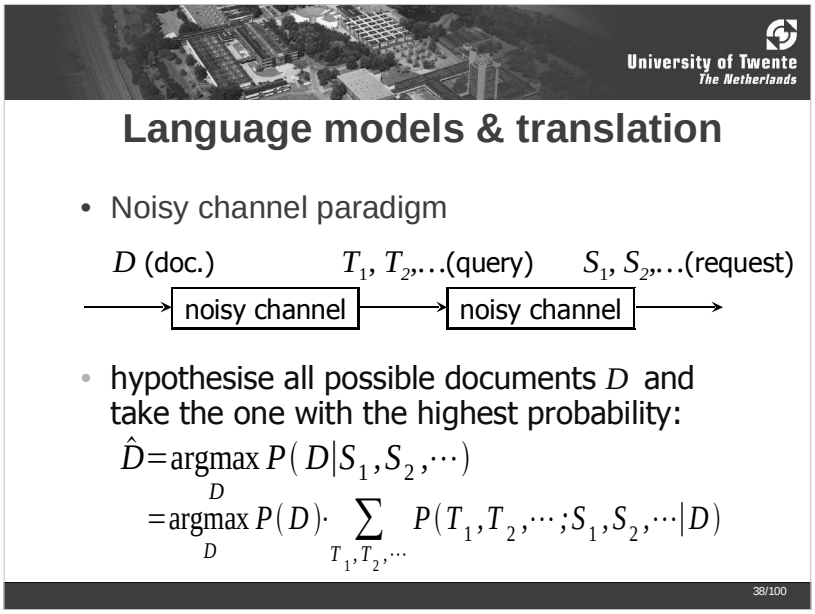
*cross-language information retrieval*
*zoeken in anderstalige informatie*
*recherche d'informations multilingues*




## Language models & translation

- Cross-language information retrieval (CLIR):
  - Enter query in one language (language of choice) and retrieve documents in one or more other languages.
  - The system takes care of automatic translation


MSN Search Result for - zoeken in anderstalige informatie - Microsoft Internet Explorer
MSN Search Result for - cross-language information retrieval - Microsoft Internet Explorer
Web Search
Stop having your access denied?
Computing Channel
Planning a trip?
Travel Channel
Search the web for:
cross-language information retrieva
Search
Yellow Pages
New Search
Cross-Language Information Retrieval Resources
Cross-Language Information Retrieval
Cross-Language Text And Speech Retrieval
Twenty-One - Cross-Language Information Retrieval links
Cross-Language Information Retrieval
Links2Go: Information Retrieval
ACM Digital Library: QUILT: implementing a large-scale cross-language text retrieval system
TWLT 14 - Language Technology in Multimedia Information Retrieval
Cross Language Information Retrieval
HCI Processor Project/Cross-Language Information Retrieval
Cross-Language Information Retrieval Resources
SIGIR'98 papers: Cross-Language Information Retrieval with the UMLS Metathesaurus
USE IT! - Unified Search Engine for Internet


University of Twente
The Netherlands
Language models & translation
Noisy channel paradigm
D (doc.)
T1, T2,…(query)
S1, S2,…(request)
noisy channel
noisy channel
hypothesise all possible documents D and take the one with the highest probability:
$\hat{D} = \arg\max_D P(D|S_1, S_2, \cdots)$
$= \arg\max_D P(D) \cdot \sum_{T_1, T_2, \cdots} P(T_1, T_2, \cdots; S_1, S_2, \cdots | D)$
38/100

University of Twente
The Netherlands

## Language models & translation

- Cross-language information retrieval :
  - Assume that the translation of a word/term does not depend on the document in which it occurs.
  - if: $S_1, S_2, \ldots, S_n$ is a Dutch query of length $n$
  - and $t_{i1}, t_{i2}, \ldots, t_{im}$ are $m$ English translations of the Dutch query term $S_i$

$$P(S_1, S_2, \ldots, S_n | D) =$$

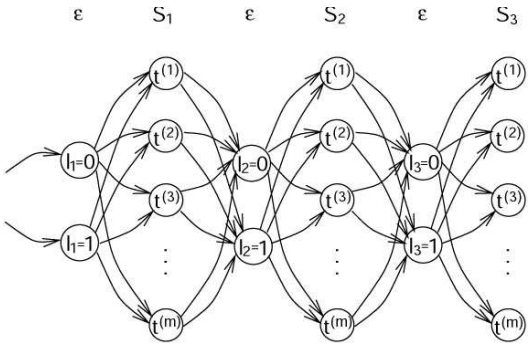
$$\prod_{i=1}^{n} \sum_{j=1}^{m_i} P(S_i | T_i = t_{ij})((1-\lambda_i) P(T_i = t_{ij}) + \lambda_i P(T_i = t_{ij} | D))$$

39/100

University of Twente
The Netherlands

## Language models & translation

- Presentation as hidden Markov model




40/100


## Language models & translation

- How does it work in practice?
  - Find for each Dutch query term $N_i$ the possible translations $t_{i1}, t_{i2},\ldots, t_{im}$ and translation probabilities
  - Combine them in a structured query
  - Process structured query




## Language models & translation

- Example:
  - Dutch query: *gevaarlijke stoffen*
  - Translations of *gevaarlijke* : *dangerous* (0.8) or *hazardous* (1.0)
  - Translations of *stoffen* : *fabric* (0.3) or *chemicals* (0.3) or *dust* (0.1)
  - Structured query:

    ((*0.8 dangerous* ∪ *1.0 hazardous*) ,

    (*0.3 fabric* ∪ *0.3 chemicals* ∪ *0.1 dust*))


## Language models & translation

- Other applications using the translation model
  - On-line stemming
  - Synonym expansion
  - Spelling correction
  - 'fuzzy' matching
  - Extended (ranked) Boolean retrieval




## Language models & translation

- Note that:
  - $\lambda_i = 1$, for all $0 \leq i \leq n$ : Boolean retrieval
  - Stemming and on-line morphological generation give exact same results:

    $P(\text{funny} \cup \text{funnies}, \text{table} \cup \text{tables} \cup \text{tabled}) =$
    $P(\texttt{funni}, \texttt{tabl})$


## Experimental Results

- translation language model
  - (source: parallel corpora)
  - average precision: 0.335 (83 % of base line)
- no translation model, using all translations:
  - average precision: 0.308 (76 % of base line)
- manual disambiguated run (take best translation)
  - average precision: 0.315 (78 % of base line)
    (Hiemstra and De Jong 1999)




## Prior probabilities

University of Twente
The Netherlands

## Prior probabilities and static ranking

- Noisy channel paradigm (Shannon 1948)

$D$ (document) → noisy channel → $T_1, T_2, \ldots$ (query)

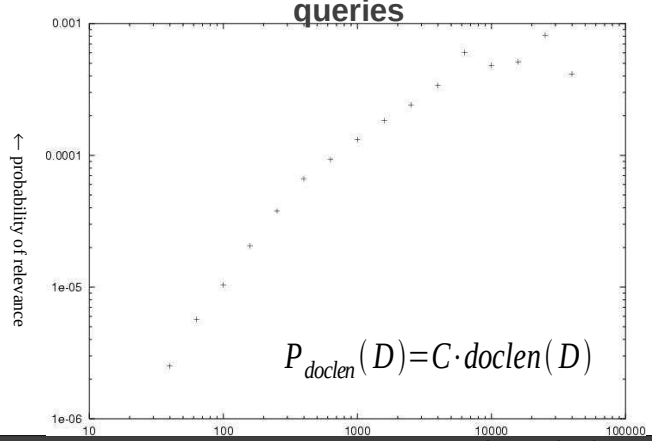
- hypothesise all possible documents $D$ and take the one with the highest probability, symbolically:

$$\hat{D} = \arg\max_D P(D|T_1, T_2, \cdots)$$
$$= \arg\max_D P(D) \cdot P(T_1, T_2, \cdots|D)$$

47/100

University of Twente
The Netherlands

## Prior probability of relevance on informational queries




48/100

University of Twente
The Netherlands

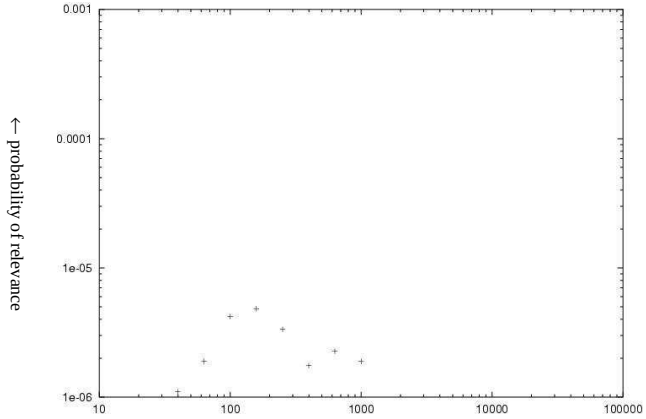
## Priors in Entry Page Search

- Sources of Information
  - Document length
  - Number of links pointing to a document
  - The depth of the URL
  - Occurrence of cue words ('welcome','home')
  - number of links in a document
  - page traffic



University of Twente
The Netherlands

Prior probability of relevance on navigational queries



University of Twente
The Netherlands

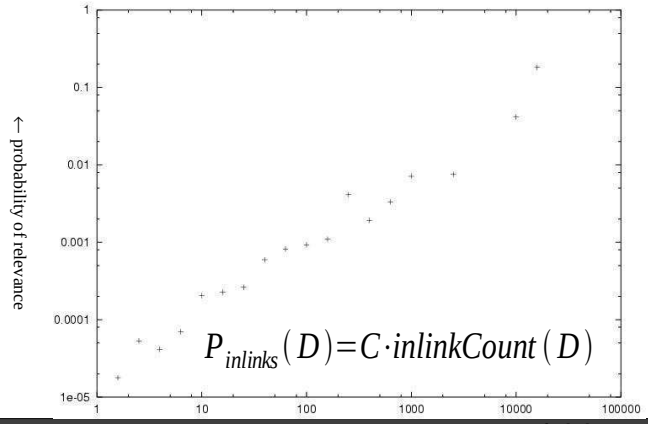
## Priors in Entry Page Search

- Assumption
  - Entry pages referenced more often
- Different types of inlinks
  - From other hosts (recommendation)
  - From same host (navigational)
- Both types point often to entry pages

51/100

University of Twente
The Netherlands

## Priors in Entry Page Search




52/100


## Priors in Entry Page Search: URL depth

- Top level documents are often entry pages
- Four types of URLs
  - root: **`www.bcs.org`**
  - subroot: **`www.bcs.org/news`**
  - path: **`www.bcs.org/news/2008`**
  - file: **`www.bcs.org/news/2008/CI.html`**




## Priors in Entry Page Search: results

| method | Content | Anchors |
|---|---|---|
| $P(Q\|D)$ | 0.3375 | 0.4188 |
| $P(Q\|D)P_{doclen}(D)$ | 0.2634 | 0.5600 |
| $P(Q\|D)P_{inlink}(D)$ | 0.4974 | 0.5365 |
| $P(Q\|D)P_{URL}(D)$ | 0.7705 | 0.6301 |

(Kraaij, Westerveld and Hiemstra 2002)


## Exercise 3 & 4 (and a break)

3. Use the following statistical translation dictionary and re-calculate your scores in the translation model:
   - $P(y1 \mid z1) = 0.5$, $P(y1 \mid z2) = 1.0$
   - $P(y2 \mid z3) = 0.2$, $P(y2 \mid z4) = 0.1$

4. Use a length prior and re-calculate scores




## Relevance models & an application to expert search


## What about relevance feedback?

- We assume that a (one) relevant document has generated the query
- So, once we find that document, we might as well stop.
- What we need is a model of "relevance", or language models of sets of relevant documents




## Lavrenko's relevance model

- "Construct a relevance model $P(T|R)$ by assuming that once we pick a relevant document *D*, the probability of observing a word is independent from the set of relevant documents"

$$P(T|R) = \sum_{D \in R} P(T|D)\,P(D|R)$$

- we only have information about *R* through a query

$$P(T|q_1,\ldots) = \sum_{D \in R} P(T|D)\,P(D|q_1,\ldots)$$


## Lavrenko's relevance model 1

- Is really a blind feedback method:
  - do an initial run and assign $P(D|q_1,...)$
  - for every retrieved document, get the most frequent terms *T*, and assign those $P(T|D)$
  - multiply both probabilities, and sum them for each document retrieved




## Balog's expert finder

- As in Lavrenko's method, use query to retrieve some initial documents.
- Instead of query (term) expansion, do person name expansion
  - for every retrieved document, get the candidates *ca*, and assign those $P(ca \mid D)$
  - multiply both probabilities, and sum them for each document retrieved

    (Balog et al. 2006)


## Balog's expert finder

- "Construct a *candidate* model $P(ca|R)$ by assuming that once we pick a relevant document *D*, the probability of observing a *candidate expert* is independent from the set of relevant documents"

$$P(ca|R)=\sum_{D\in R} P(ca|D)P(D|R)$$

- we only have information about *R* through a query

$$P(ca|q_1,...)=\sum_{D\in R} P(ca|D)P(D|q_1,...)$$

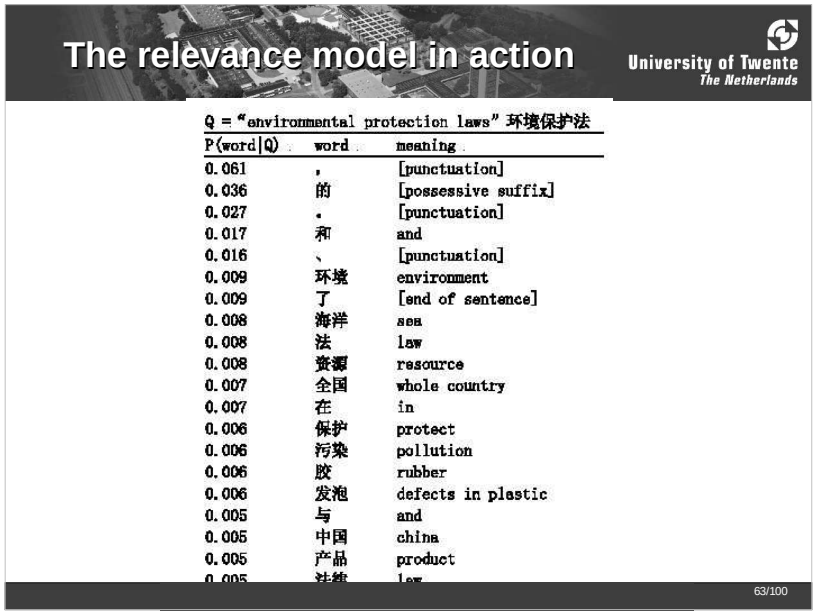
$$\sum_{D\in R} P(ca|D)\prod_{i=1}^{n}((1-\lambda)P(q_i)+\lambda P(q_i|D))$$

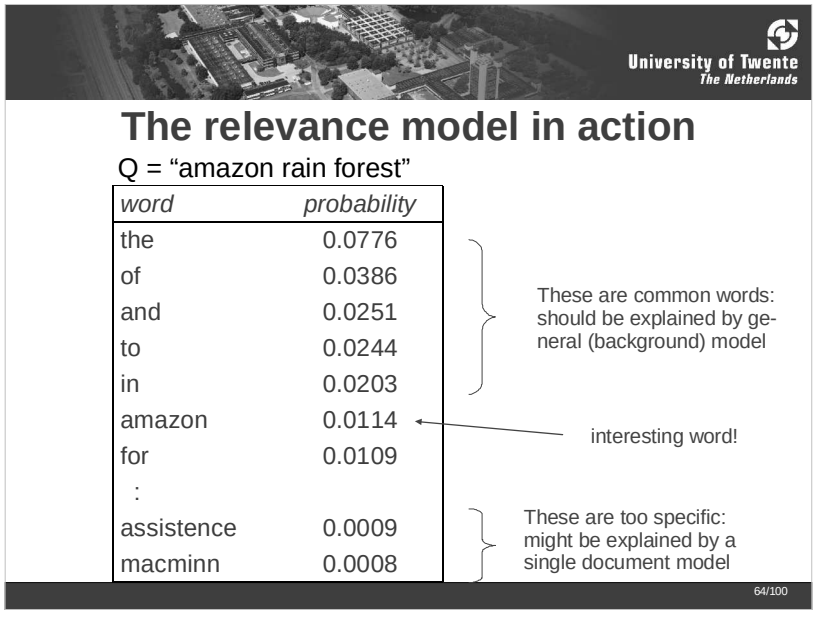



## Balog's expert finder

| | #rel | MAP | R-prec | MRR | P10 |
|---|---|---|---|---|---|
| Model 1 (candidate model): | | | | | |
| BASE | **511** | **0.1253** | **0.1914** | 0.2759 | **0.236** |
| Model 2 (document model): | | | | | |
| BASE | **580** | **0.1880** | 0.2332 | 0.5149 | **0.316** |

Figure 2, Candidate model vs. document model


The relevance model in action
University of Twente
The Netherlands
Q = "environmental protection laws" 环境保护法
P(word|Q) word meaning
0.061 , [punctuation]
0.036 的 [possessive suffix]
0.027 . [punctuation]
0.017 和 and
0.016 、 [punctuation]
0.009 环境 environment
0.009 了 [end of sentence]
0.008 海洋 sea
0.008 法 law
0.008 资源 resource
0.007 全国 whole country
0.007 在 in
0.006 保护 protect
0.006 污染 pollution
0.006 胶 rubber
0.006 发泡 defects in plastic
0.005 与 and
0.005 中国 china
0.005 产品 product
63/100


University of Twente
The Netherlands
The relevance model in action
Q = "amazon rain forest"
word probability
the 0.0776
of 0.0386
and 0.0251
to 0.0244
in 0.0203
amazon 0.0114
for 0.0109
:
assistence 0.0009
macminn 0.0008
These are common words: should be explained by general (background) model
interesting word!
These are too specific: might be explained by a single document model
64/100


## What we need is *parsimony*

- Optimize the probability to predict language use
- Minimize the total number of parameters needed for that
- Expectation Maximization Training
  (Hiemstra, Robertson and Zaragoza 2004).

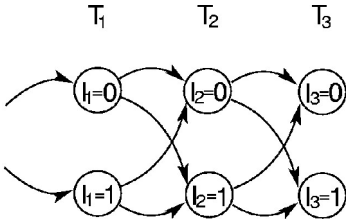



Expectation Maximization Training
&
An application of expert search


## Statistical language models

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

- Presentation as hidden Markov model
  - finite state machine: probabilities governing transitions
  - sequence of state transitions cannot be determined from sequence of output symbols (i.e. are hidden)




67/100


## Fundamental questions for HMMs

1. Given a model, how do we efficiently compute the probability $P(O)$ of the observation sequence $O$ ?
2. Given the observation sequence $O$ and a model how do we choose a state sequence that best explains the observations?
3. Given an observation sequence $O$ how do we find the model that maximises the probability $P(O)$ of the observation sequence $O$ ?

68/100


## Fundamental answers

1. Forward procedure or backward procedure
2. Viterbi algorithm
3. Baum Welch algorithm / forward-backward algorithm (special case of the expectation maximisation-algorithm, or "EM-algorithm")




## Statistical language models

- Re-estimate the value of $\lambda$ from relevant documents (relevance feedback)
  - Expectation Maximisation algorithm
  - Estimate different value of $\lambda_i$ for each term (i.e. different importance of each term.)


## Parsimonious models

- Define background models, document models and relevance models in a layered fashion
  1. First define background model
  2. Higher order model(s) should not model language that is well explained by the background model already
  3. Use EM training (we'll see how later on)




## How does it work?

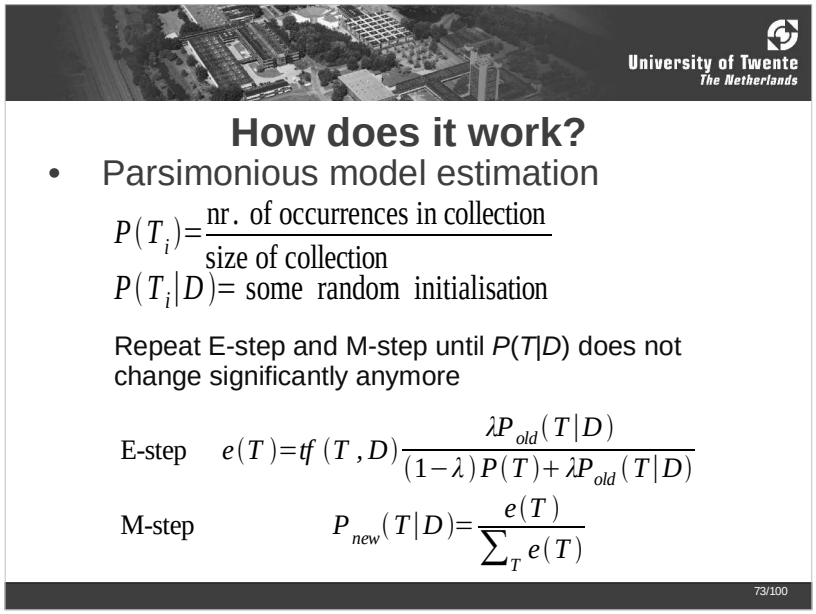
- Remember this equation?

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

- In the old days:

$$P(T_i) = \frac{\text{nr. of occurrences in collection}}{\text{size of collection}}$$

$$P(T_i|D) = \frac{\text{nr. of occurrences in document}}{\text{size of document}}$$




## How does it work?

- Parsimonious model estimation

$$P(T_i) = \frac{\text{nr. of occurrences in collection}}{\text{size of collection}}$$

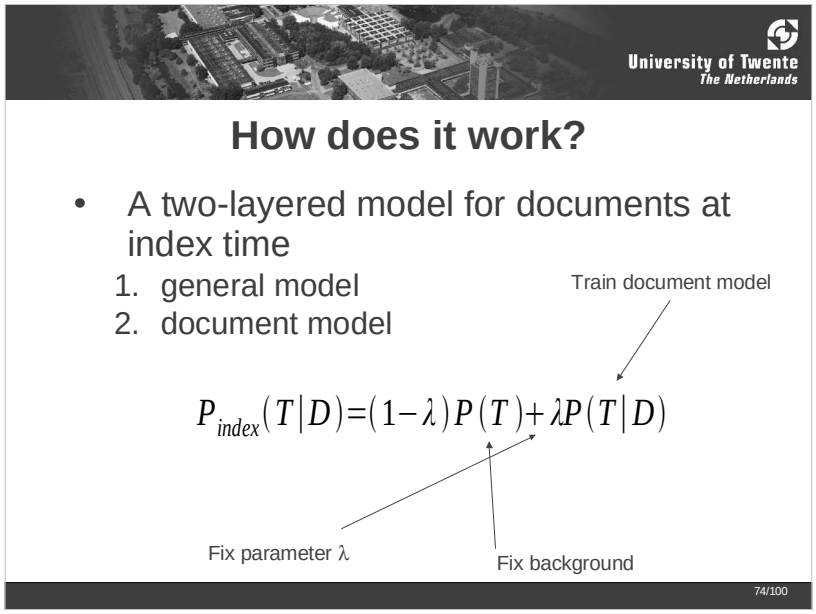
$$P(T_i|D) = \text{some random initialisation}$$

Repeat E-step and M-step until $P(T|D)$ does not change significantly anymore

E-step $$e(T) = tf(T,D)\frac{\lambda P_{old}(T|D)}{(1-\lambda)P(T)+\lambda P_{old}(T|D)}$$

M-step $$P_{new}(T|D) = \frac{e(T)}{\sum_T e(T)}$$






## How does it work?

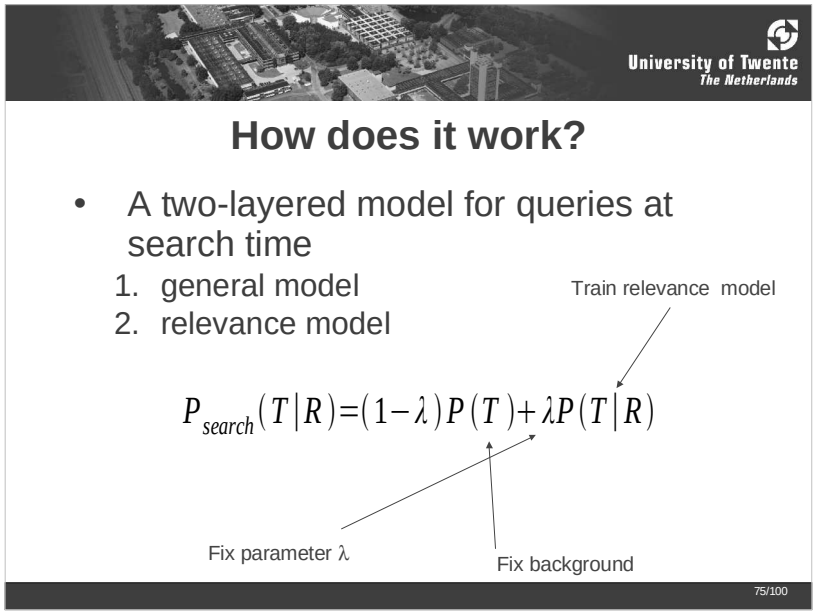
- A two-layered model for documents at index time
  1. general model
  2. document model

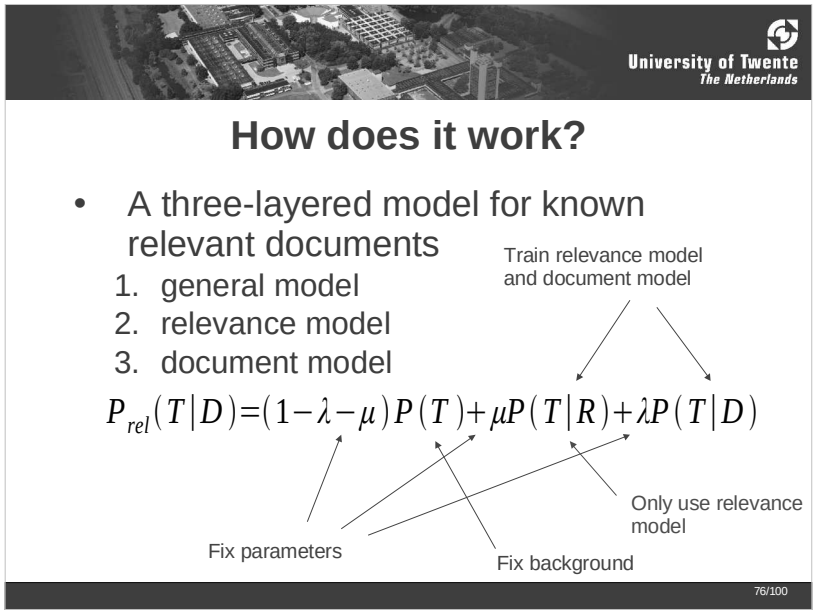
Train document model

$$P_{index}(T|D) = (1-\lambda)P(T)+\lambda P(T|D)$$

Fix parameter λ

Fix background


University of Twente
The Netherlands
How does it work?
A two-layered model for queries at search time
1. general model
2. relevance model
Train relevance model
$P_{search}(T|R)=(1-\lambda)P(T)+\lambda P(T|R)$
Fix parameter λ
Fix background
75/100


University of Twente
The Netherlands
How does it work?
A three-layered model for known relevant documents
1. general model
2. relevance model
3. document model
Train relevance model and document model
$P_{rel}(T|D)=(1-\lambda-\mu)P(T)+\mu P(T|R)+\lambda P(T|D)$
Only use relevance model
Fix parameters
Fix background
76/100

University of Twente
The Netherlands

## How to use a relevance model?

- Measure cross-entropy between relevance model and document model

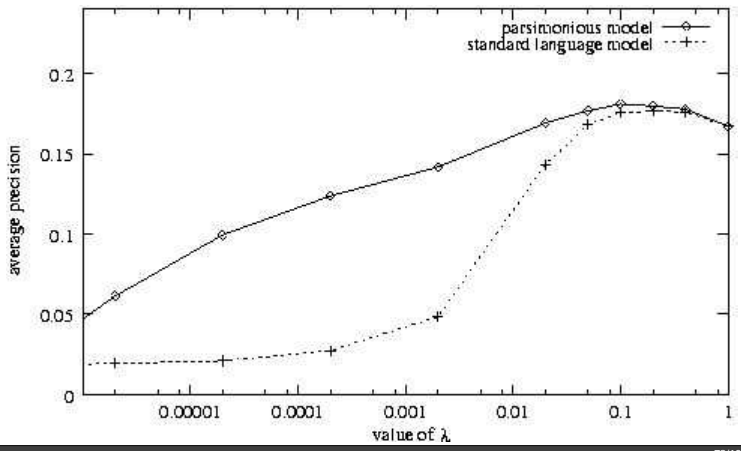
$$H(R,D) = -\sum_{T} P(T|R) \log((1-\lambda)P(T) + \lambda P(T|D))$$

only terms with non-zero $P(T|R)$ contribute to sum

77/100

University of Twente
The Netherlands

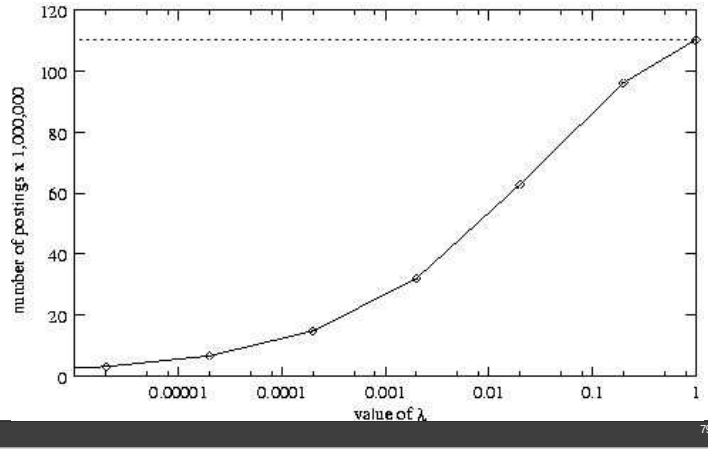
## So, what happens?




78/100

University of Twente
The Netherlands

## How much are we throwing away?




79/100

University of Twente
The Netherlands

## "amazon rain forest" again

$\mu = 0.0000001$

| *word* | *probability* |
|---|---|
| amazon | 0.3367 |
| rain | 0.3365 |
| forest | 0.2896 |
| ban | 0.0370 |
| brazil | 0.0002 |

80/100

University of Twente
The Netherlands

## Serdyukov's expert model

- Use an email archive to search for experts
- Experts both send and receive email on the topic they know well
- Each email is a mixture of the language models of each potential expert
  - i.e. because of in-line quotations

$$P_{rel}(T|D) = \sum_{e \in D} P(T|E=e)P(E=e|D)$$

Train expert models

Fix parameters

81/100

University of Twente
The Netherlands

## Serdyukov's expert model

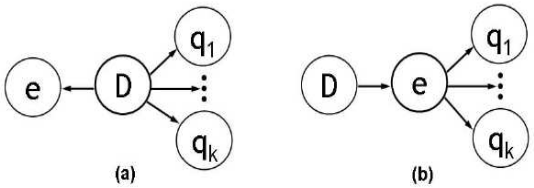



Fig. 1. Dependence networks for two methods of estimating $P(e, q_1, ..., q_k)$

82/100


## EM-training for expert search

| Method | MAP | MRR | R-prec | P5 | P10 | P20 |
|---|---|---|---|---|---|---|
| Method 1 | 0.1587 | 0.6550 | 0.2598 | 0.4285 | 0.4122 | 0.3341 |
| Method 2 | 0.1712 | 0.6712 | 0.2755 | 0.4306 | 0.4304 | 0.3653 |

**Table 1: Performance of expert ranking methods**

- (Serdyukov and Hiemstra 2008)
  (table contains results from earlier experiments)




## Probabilistic Latent Semantic Indexing

- Each document is a mixture of a number of latent models (or topics)
- We do not know what document discusses what topics

$$P_{rel}(T|D) = \sum_{m} P(T|M=m)\,P(M=m|D)$$

Only fix the number of models

Train latent models

Train model weights


## Probabilistic Latent Semantic Indexing

- Related to Singular Value Decomposition
- Problems with over-training
  (Hofmann 1999)




## Exercise 5 & 6

5. Find the experts for the query *y* using Balog's expert finder model, using *only* your document
   – Take a uniform $P(ca|D)$ in each document

6. Think about how you would do the EM-training of Serdyukov's expert finder model
   5 minutes!


Probabilistic Random Walks
&
An application to expert search




## Approach towards entity ranking

1. off-line preparation: index corpus with entity tagging. use NLP techniques to recognize entities if the are not tagged.
2. on-line, query dependent: building of an entity containment graph from top ranked retrieved documents
3. relevance propagation within the graph and output entities of interest in order of their relevance.


## NLP tagging

- XML fragment

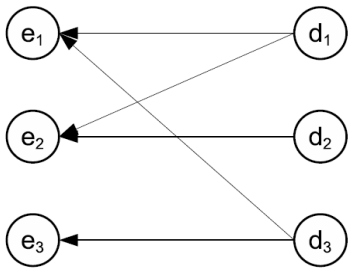
<entry> Jorge Castillo (artist) Castillo greatly admired Pablo Picasso, and that influence shows his paintings, etchings, and lithographs ...

- tagged fragment

<entry><s><enamex.person>Jorge Castillo</enamex.person> <O.PUNC>(</O.PUNC> <O.NN>artist</O.NN><O.PUNC>) </O.PUNC> </s><s><enamex.person>Castillo</enamex.person> <O.RB>greatly</O.RB> <O.VBD>admired</O.VBD> <enamex.person>Pablo Picasso</enamex.person><O.PUNC>, </O.PUNC> <O.CC>and</O.CC><O.DT>that</O.DT> <O.NN>influence</O.NN> <O.VBZ>shows</O.VBZ><O.IN >in</ O.IN> <O.PRPDOLAR>his</O.PRPDOLAR> ...




## Including Further Entity Types

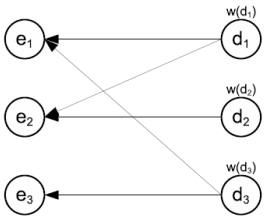


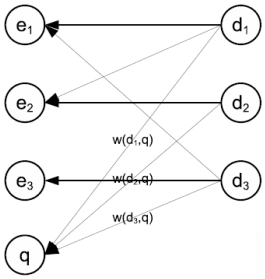

- We model with entity containment graphs the relationship between entities and documents.
- Documents and Entities are represented as vertices.
- Edges symbolize the containment relation.

University of Twente
The Netherlands

## Modelling query-dependent scores

- Model 1: vertex weights
- Model 2: additional query node and edge weights

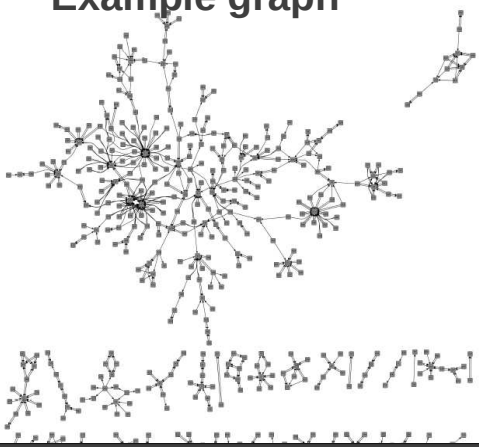




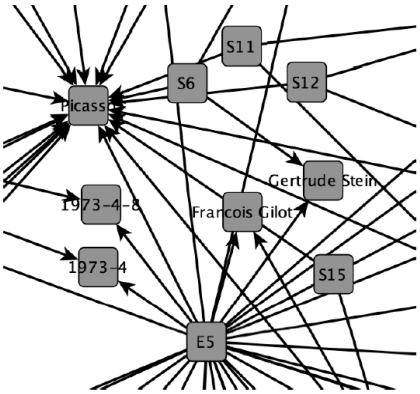

University of Twente
The Netherlands

## Example graph



University of Twente
The Netherlands

## Entity identity




- identity check: Is Gilot the same person as Francois Gilot?
- precision: How do we model the occurrence of April 8, 1973 and 1973?

93/100

University of Twente
The Netherlands

## Probabilistic random walk

- The mutually recursive definition describes a walk over the different type of edges in the graph: query–doc, doc–doc, doc–ent, ent–ent.

**Probabilistic Random Walk**

$$P(e) = \lambda_1 \sum_d P(e|d)P(d) + \lambda_2 \sum_{e'} P(e|e')P(e')$$

$$P(d) = \lambda_0 P(d|q) + \lambda_1 \sum_e P(d|e)P(e) + \lambda_2 \sum_{d'} P(d|d')P(d')$$

94/100


## Experimental Results

Performance overview of the relevance propagation models:

| *Model* | *unweighted* | *weighted* |
|---|---|---|
| MAX | | 0.352 |
| IDG | 0.342 | 0.371 |
| HITS | 0.343 | 0.376 |
| PRW | 0.340 | 0.386 |

(Rode et al. 2007)




## Exercise 7

- Draw graph model 2 (with query node) for the query *y*
  - Take initially zero probability of nodes, except for the query node which get 1
  - Do two relevance propagation steps


## Advanced models conclusion

- Translation model: accounts for multiple query representations (e.g. CLIR or stemming) (*exercise 3*)
- Document priors: account for "non-content" information (*exercise 4*)
- Relevance models: query expansion using initial ranked list (*expert search exercise 5*)
- Expectation Maximization Training: estimate the probability of unseen events (*expert search exercise 6*)
- Random walks: find most central entity/document (*expert search exercise 7*)





- Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. Formal models for expert nding in enterprise corpora. *Proceedings of SIGIR 2006*
- 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: Research and Advanced Technology for Digital Libraries, Springer-Verlag*, 1998
- Djoerd Hiemstra and Franciska de Jong. Disambiguation strategies for cross-language information retrieval., *Lecture Notes in Computer Science 1696: Research and Advanced Technology for Digital Libraries, Springer-Verlag*, 1999
- Djoerd Hiemstra, Stephen Robertson and Hugo Zaragoza. Parsimonious Language Models for Information Retrieval", In *Proceedings of SIGIR 2004*

## References

University of Twente
The Netherlands


- Thomas Hofmann, Probabilistic latent semantic indexing, *Proceedings of SIGIR 1999*
- Wessel Kraaij, Thijs Westerveld and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. In *Proceedings of SIGIR 2002*
- Victor Lavrenko and Bruce Croft. Relevance based language models. *Proceedings of SIGIR 2001*
- Stephen Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*(3):129–146, 1976
- Henning Rode, Pavel Serdyukov, Djoerd Hiemstra, and Hugo Zaragoza, "Entity Ranking on Graphs: Studies on Expert Finding'', Technical Report 07-81, CTIT, 2007
- Pavel Serdyukov and Djoerd Hiemstra, Modeling documents as mixtures of persons for expert finding, In *Proceedings of ECIR* 2008
- Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal 27*, 1948




University of Twente
The Netherlands

## Acknowledgments


- Some slides were kindly provided by:
  - Pavel Serdyukov
  - Henning Rode