IR4IP Tutorial

IPI Confex, March 2009

Erik Graf and Thomas Roelleke

Introduction

IPI Confex, March 2009

Thomas Roelleke
Queen Mary University of London

Outline

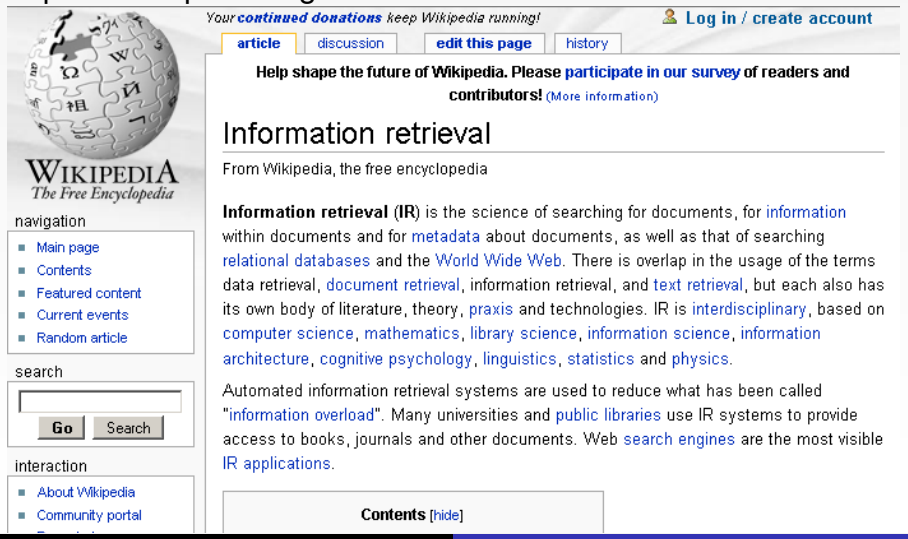
Structure

1200 - 1330: 90 mins, “six” 15 mins slots.

1. Introduction (TR): 15 mins
2. Indexing (EG): 15 mins
3. Retrieval Models (TR): 15 mins
4. Interaction (EG): 15 mins
5. Evaluation (TR): 15 mins

What does Wikipedia say?

http://en.wikipedia.org/wiki/Information_retrieval



10 Issues in IR

1. Retrieval models (ranking functions)
2. Text processing ("Indexing"): NLP / understanding
3. Interactivity
4. Efficiency: Compression, parallel IR
5. Distributed IR (data fusion, meta retrieval)
6. Multimedia: image, video, sound
7. Evaluation
8. Web retrieval (link analysis)
9. Cross-lingual IR
10. Digital libraries: IR application

see http://www.dlib.org/dlib/november95/11croft.html

IR Herbstschule 2008

IR for PhD and post-doc researchers.
http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=08402

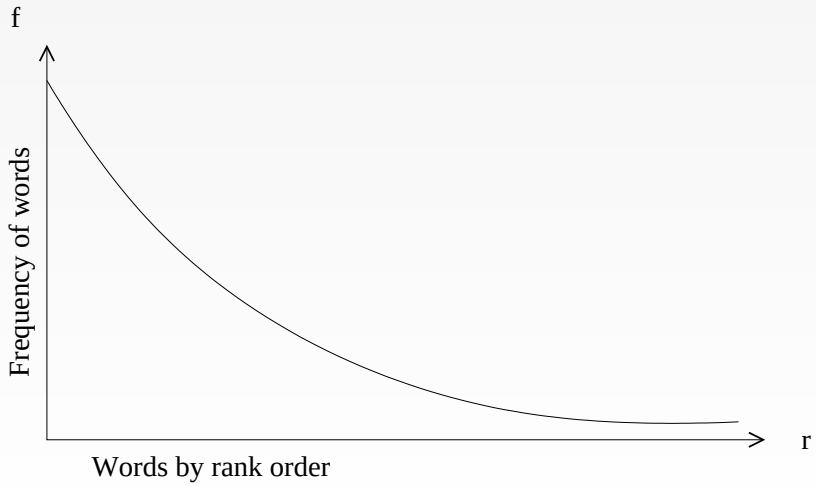
Empty Answer and Many Answer "Problem"

- "Empty Answer Problem":
  - Free-text: Find a web page that offers boats that are fast AND comfortable AND NOT EXPENSIVE
  - SQL DB: SELECT * FROM properties WHERE price $<$ 200k AND bedrooms $>$ 3 AND location LIKE 'London';

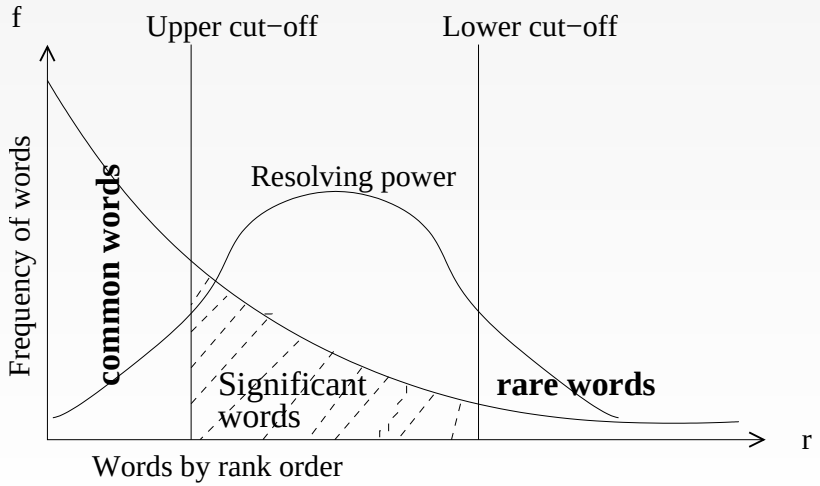
  Query is too *narrow*, too *specific*.
- "Many Answer Problem":
  - SELECT * FROM properties WHERE price $<$ 200k OR bedrooms $>$ 3 OR location LIKE 'London';

  Query is too *general*, too *exhaustive*.

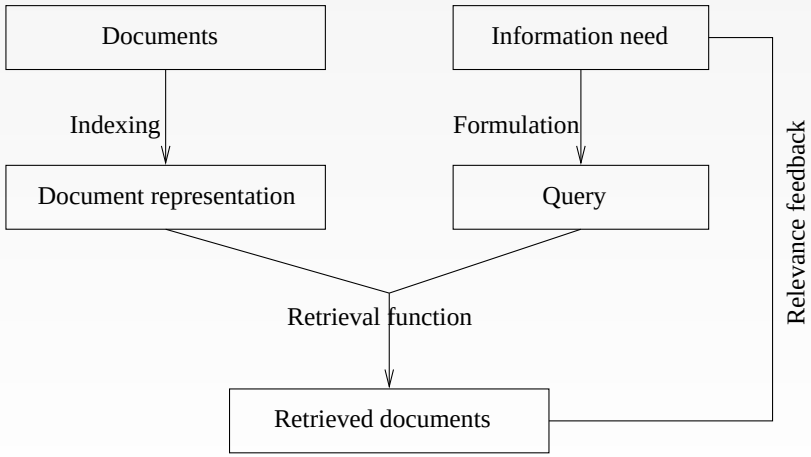
Zipf's law



Luhn's analysis



Conceptual Model for IR



The 60s/70s: VSM and "The" probabilistic retrieval model

- Boolean retrieval "semiconductors AND inventor=ibm" still popular.
- VSM: Vector-space model: Replace Boolean retrieval by retrieval based on the distance of document and query vectors
- Probabilistic justification of what ranking is? $P(r|d,q)$, the probability of relevance.
  Illustration: Present a document-query pair to several users, and each user assesses the relevance.
  After ... steps: A term/keyword is GOOD, if it occurs more often in relevant documents than in non-relevant documents.

The 80s

- SQL databases become widely available ... text processing?
- VSM, probability of relevance, all interesting, but not good enough?
- Therefore, view IR as "logical implication" ... Rijsbergen's $P(d \rightarrow q)$

Mid 90s: The web

- The web. Keyword-based (content-based) retrieval alone not good enough.
- Pagerank ... Google ... A page is good if it is "popular".
- A popular page is an authority?
- Combine content-based (keyword-based) ranking with pagerank.
- The page rank (popularity, authority) is independent of the query!

Late 90s: Language modelling

- VSM, probabilistic model, logical model, all interesting, but how about language modelling?
- Given a document and a collection, there is a probability that the query is generated from those *TWO* "models"!

Web/page search OR Enterprise search OR Semantic web?

- Web search: surface/horizontal search
- Enterprise search: "deep" search, semantic search, vertical search
- Semantic web, webDB: explore/exploit the web similar to what is known for enterprise search

Summary

- Information retrieval: not just "document" retrieval
- An information need is translated into a query: "loss of information"
- Hypertext/Web: Content (words/terms) + Links (links, popularity, authority)

Questions?

Thank you!

IR4IP Tutorial

IPI Confex, March 2009

Erik Graf and Thomas Roelleke

Indexing

IPI Confex, March 2009

Erik Graf
University of Glasgow

Outline

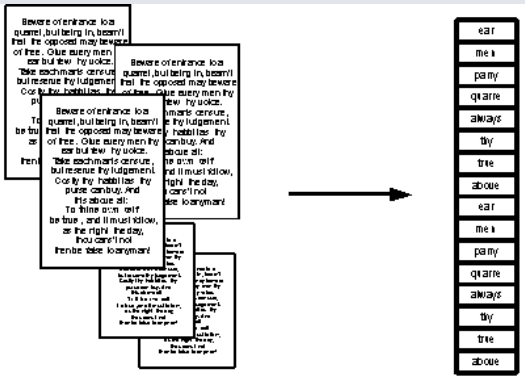
Aim

Storage of information in a way that supports efficient retrieval.
Two main points of considerations:

- Accuracy of semantic representation.
- Space and time efficiency.

Overview of Indexing Process

Basic concept

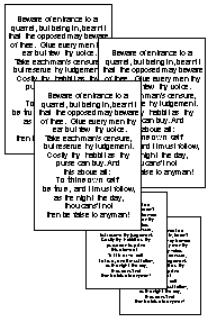


Document Collection

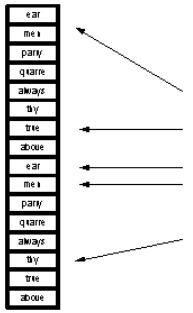
Index

Overview of Indexing Process

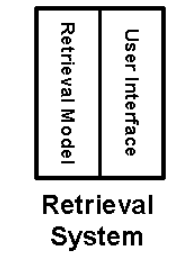
Basic concept



Document Collection



Index



Retrieval System

From UniTerm to Full text

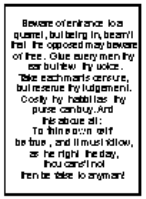
Index of books

Document Collection

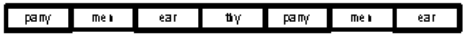
Index

From UniTerm to Full text

Fulltext indices



Document

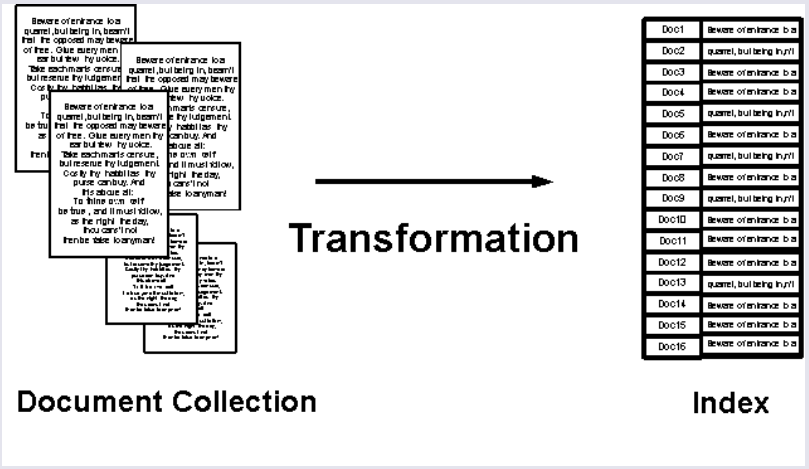


Representation

- Represent documents via the complete set of terms.

Index creation

Process overview

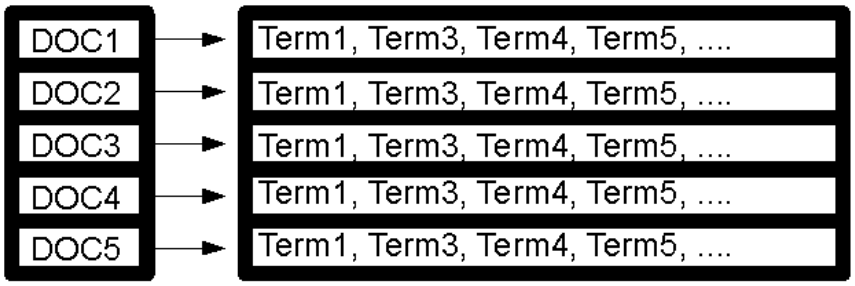


Index creation

Term Processing

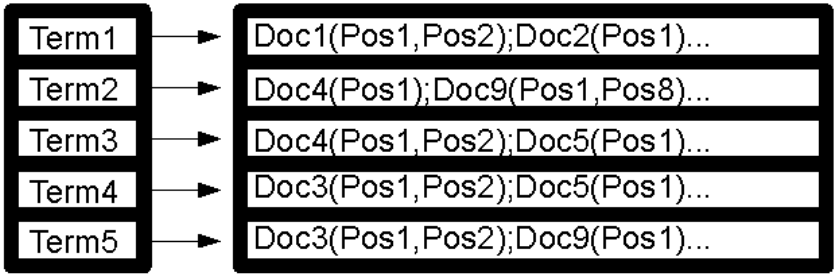
- **Tokenization**: Extract the terms from a document. Removal of tags, punctuation.
- **Stop-wording**: Remove terms with very high document frequencies (e.g. ‘of’, ‘the’).
  In the English language these are responsible for approx. 30% of term volume.
  Stop word lists are getting shorter; or often no stop-wording is applied.
- **Stemming**: Collapsing morphological variants of words (i.e ‘Patent’, ‘Patents’, ‘patenting’ is stemmed to ’patent’). Eases querying and reduces index size.

Direct Index



- Optimized file system for retrieval. Mimics a Unix ’grep’ or Windows98 search.
- Scales badly with respect to number of documents.

Inverted Index



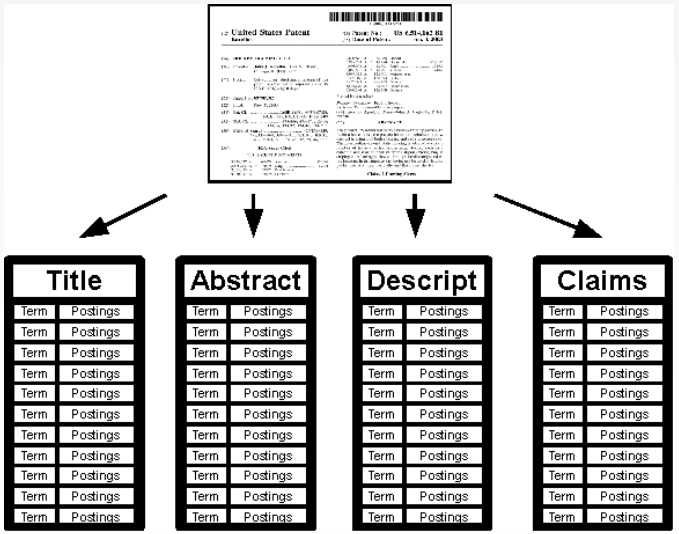
- Default index structure in Information Retrieval.
- Computationally very efficient. Scales well.

Positional Index

Positional Index

- Stores the position of term in addition to the posting and frequency
- Allows for word-order- ,phrasal- , offset-querying.
- Increases the size of the index by 2-10 times (depending on the average document length)

Field Index



Field Index

- Enables part specific searching
- Allows assigning weight to different parts or aspects of documents (Robertson BM25F)
- A form of partitioning documents.

Virtual fields

Virtual fields

- Enrich document with document-external data
- Document priors (Web: Number of inlinks, URL length, URL text)
- Another example anchor text:
  - Link: ‘http://www.acm.org/sigir/’
  - Text: ‘ACM SIGIR: Information Retrieval Special Interest Group’

Summary

'Trends' in document representations

- Representations resemble documents more closely
  - Full text
  - Positional
- Exploitation of semantic 'hints'
  - Structure
  - Style
- Enrichment of document with meta-data
  - Anchor text
  - Document priors (inlinks,URL, etc ...)

Questions?

Thank you!

IR4IP Tutorial

IPI Confex, March 2009

Erik Graf and Thomas Roelleke

Retrieval Models

IPI Confex, March 2009

Thomas Roelleke
Queen Mary University of London

Outline

Task of an IR model

- Process a query such that the result is specific (few hits only, and hits on topic) while being exhaustive (enough hits, good coverage).
- Retrieve relevant documents while not retrieving non-relevant documents.
- Rank documents.

Time frame

60s/70s: Vector-space model (VSM); TF-IDF (term frequency — inverse document frequency) probabilistic (binary independence) retrieval model (BIR)

Early/mid 90s: Poisson, BM25 (best-match version 25)

Mid 90s: Page-rank (Authority-based retrieval a-la Google)

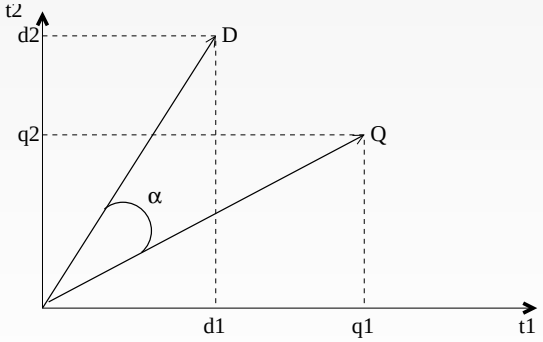
Late 90s: Language Modelling

Term	DocId
sailing	doc1
boats	doc1
sailing	doc2
boats	doc2
sailing	doc2
sailing	doc3
east	doc3
coast	doc3
sailing	doc4
boats	doc5

$\mathrm{tf}(t,d)$	term frequency
$\mathrm{df}(t)$	document frequency
N_D	number of documents

$$\begin{aligned} \mathrm{tf}(\text{sailing}, \text{doc2}) &= 2 \\ \mathrm{df}(\text{sailing}) &= 4 \\ N_D &= 5 \\ \mathrm{idf}(\text{sailing}) &= -\log \frac{4}{5} = \ldots \end{aligned}$$

Vector-Space Model (VSM)



$$\mathrm{RSV}_{\mathrm{VSM}}(d,q) := \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \cdot |\vec{q}|}$$

TF-IDF

TF: Term Frequency of TERM in DOCUMENT
IDF: Inverse Document Frequency of TERM in COLLECTION

$$\begin{aligned} \mathrm{tf}(t,d) &:= \text{count term occurrence within DOCUMENT} \\ \mathrm{idf}(t) &:= \text{count documents in COLLECTION} \end{aligned}$$

IDF of frequent term is small, IDF of rare term is large.
Reflects the searcher trying to find terms that are rare/discriminative overall, but frequent in the requested document.

Binary Independence Retrieval (BIR) Model

- Established mid/end 70s ([Robertson and Sparck Jones, 1976, Croft and Harper, 1979])
- Terms are “good”: if they are frequent in relevant documents and rare in non-relevant documents
- Terms are “poor”: if they are frequent in non-relevant documents and rare in relevant documents
- Famous formula (BIR term weight):

$$\text{bir}(t, r, \bar{r}) := \frac{P(t|r)}{P(t|\bar{r})}$$

Poisson Model

- Probability that an event occurs k times given that in average it occurs λ times
- Example: Probability that 4 of 7 days are sunny, knowing that in average every second day was sunny in the past (one month: 15/30, 10 years: 1800/3600).

BM25/Okapi Formula

- established since early/mid 90s
- combines a special TF component with the BIR term weight
- considers document length
- is mathematically

$$\sum_{t \in d \cap q} \text{bir}(t, r, \bar{r}) \cdot \frac{\text{tf}_d}{\text{tf}_d + k_1} \cdot \frac{\text{tf}_q}{\text{tf}_q + k_3} + k_2 \cdot \text{ql} \cdot \frac{\text{avgdl} - \text{dl}}{\text{avgdl} + \text{dl}}$$

k_1: pivoted document length ...

Language Modelling (LM)

- THE alternative to BM25/TF-IDF?
- Established late 90s
- Derived from $P(q|d)$, i.e. the probability that document d and the background model (the collection) generate the query.
  Example: ...

Classification-oriented Models

- Bayesian classifier
- Support-vector machine (SVM)
- Duality to ad-hoc retrieval: Retrieve classes for an incoming (new) document

Web/Link-based Retrieval Models

- Idea: A page that is referenced by many “good” pages is a “good” page. Note the recursive usage of “good” ([Brin and Page, 1998]).
- Authority principle.
- Apparently the break-through for Google late 90s.
- TF boosting: Propagate the anchor text terms to the referenced object; multimedia/image retrieval.

Relationships between Retrieval Models

- The content-oriented models (TF-IDF, BM25, LM) are combined with link-based models (e.g. term propagation).
- All retrieval models try to optimise the ranking.
- Can one know in advance which model is best for which query?
- Is a combination of models useful?
- Can a system learn a model? Learn when to use which model?

Summary

- Retrieval models define the ranking (scores) of retrieved objects
- Several strands of models with many ways of estimating parameters stimulate IR research

Questions?

Thank you!


Brin, S. and Page, L. (1998).
The anatomy of a large-scale hypertextual web search engine.
In *7th International WWW Conference, Brisbane, Australia.*


Croft, W. and Harper, D. (1979).
Using probabilistic models of document retrieval without relevance information.
Journal of Documentation, 35:285–295.


Robertson, S. and Sparck Jones, K. (1976).
Relevance weighting of search terms.
Journal of the American Society for Information Science, 27:129–146.

IR4IP Tutorial

IPI Confex, March 2009

Erik Graf and Thomas Roelleke

User Interaction

IPI Confex, March 2009

Erik Graf
University of Glasgow

Outline

Aim

Improving the retrieval process by exploring a user's interaction with a system

Models of Information Access

Standard Model of the Process

1. Information Need.
2. System and collection selection
3. Query formulation
4. Query submission
5. Receival of results in form of information items.
6. Scanning, evaluation, and interpretation of results.
7. Stop or recursion.

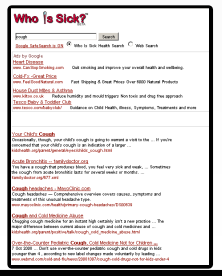
Models of Information Access

Search Interface

Query transformation

Retrieval Model

Retrieval System



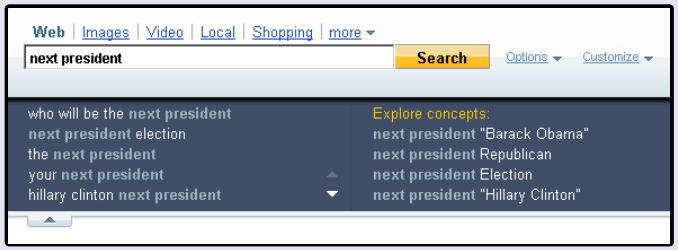
Query Specification

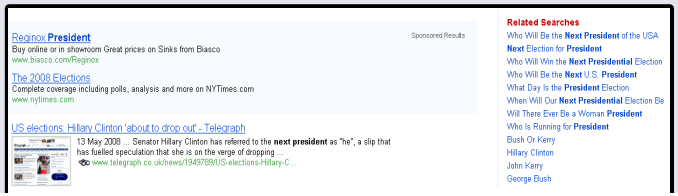
The Problem

- Synonymity, Vocabulary mismatch (car, automobile, sedan, van)
- Polysemy (Chip, Java, etc ...)
- Vague and short user queries (2-3 terms average length for web queries)

Query Specification

Keyword Suggestion





Query Specification

Keyword Suggestion Techniques

- Thesauri
- Co-Occurrence analysis of terms
- Pseudo-Relevance Feedback
- Query Clustering

Query Specification

Query Clustering

No.	Query Text	Clicked Documents
1.	law of thermodynamics	ID: 761571911 Title: Thermodynamics ID: 761571262 Title: Conservation Laws
2.	conservation laws	ID: 761571262 Title: Conservation Laws ID: 761571911 Title: Thermodynamics
3.	Newton law	ID: 761573959 Title: Newton, Sir Isaac ID: 761573872 Title: Ballistics
4.	Newton law	ID: 761556906 Title: Mechanics ID: 761556362 Title: Gravitation

Query Transformation

Automatic query expansion

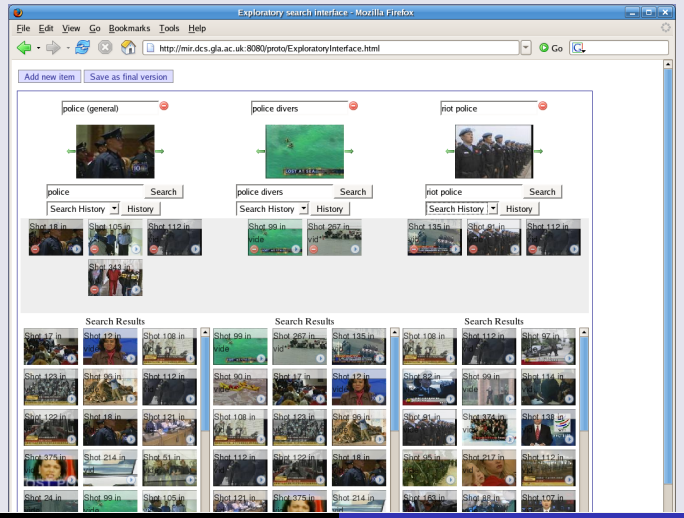
Users are generally reluctant to use suggested keywords. Therefore search engines often apply automated query expansion based on the following techniques:

- Thesauri, Co-occurrence analysis
- Pseudo-Relevance Feedback
- Query Clustering

Information item presentation

- Summarization: Which snippet will allow for the best relevance judgment from the users side.
- Result page design: Tabbed result pages, amount of shown results.

Information item presentation



Automatically derived visualizations

Clustering

Grouping of 'similar' items based on document to document similarity models.

- Hierarchical clustering (agglomerative, divisive)
- Partitional clustering

Automatically derived visualizations

Clustering

User data

Forms

- Implicit feedback: Automatically collected records of a users interaction with a system (e.g. 'click-through' data).
- Explicit feedback.

Implicitly collected records

Applications

- Meta-data for the collection documents (a query-log field index)
- Collaborative filtering (a la Amazon: people who entered this query also clicked on ...)
- Query Clustering (Keyword suggestion)
- Personalization (Personal search history, topical preferences)

Explicitly collected records

Applications

- Personalization
- Evaluation of retrieval performance.

Questions?

Thank you!

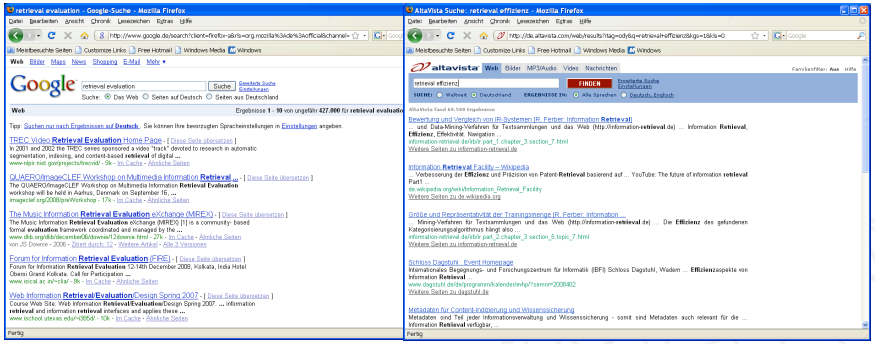
Thomas Mandl
Informationswissenschaft
Universität Hildesheim
mandl@uni-hildesheim.de

IR4IP: Evaluation

Slides by Thomas Mandl, IR Herbstschule 2008

Updates by Thomas Roelleke

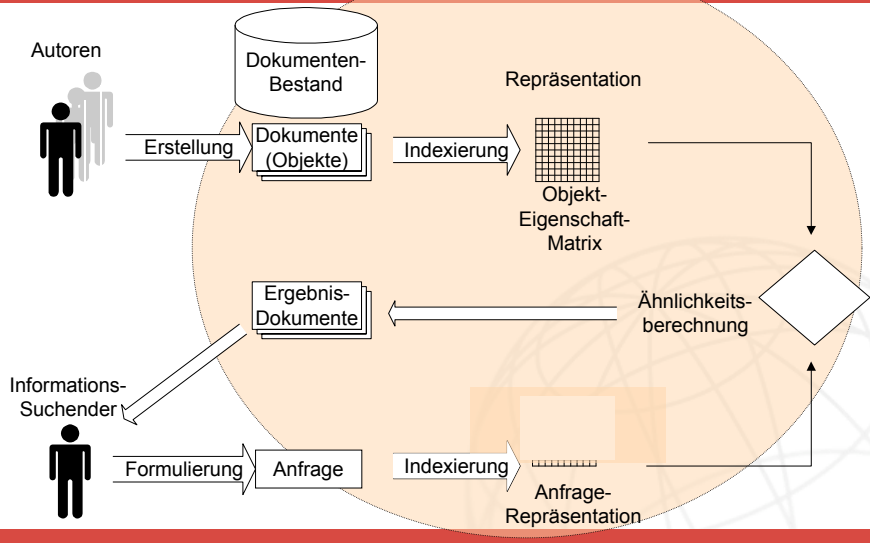
Die Evaluierung von Information Retrieval Systemen



- Which System is better?

„There must be some fundamental understanding of what it means to be good and what it means to be better“ (Bollmann/Cherniavsky 1983,3)

Many Components



Role of Evaluation

- Many components, models and optimisation techniques involved in a search system
- Effectiveness for a given (new) set of data difficult to forecast
- A general superiority of a single model or a single component is difficult to establish
- Therefore, evaluate effectiveness
- A holistic evaluation is difficult
- Measure success/satisfaction of users?

Recall und Precision

- „The ability of the retrieval system to uncover relevant documents is known as the recall power of the system“ (Lancaster 1968,55)

$$\text{Recall} = \frac{\textit{Number of retrieved relevant documents}}{\textit{Number of relevant documents}}$$

$$\text{Precision} = \frac{\textit{Number of retrieved relevant documents}}{\textit{Number of retrieved documents}}$$

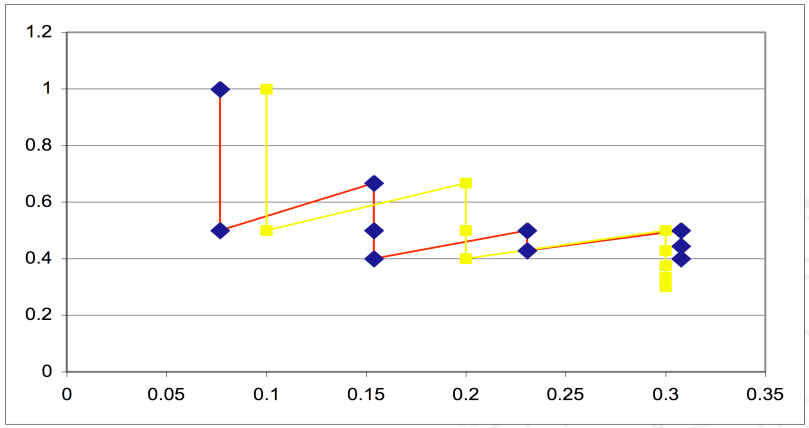
Example

- Assume that for a given query, the following documents are relevant (10 relevant documents)

  {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

- Suppose that the following documents are retrieved for that query:

rank	doc	precision	recall	rank	doc	precision	recall
1	**d123**	1/1	1/10	8	d129		
2	d84			9	d187		
3	**d56**	2/3	2/10	10	**d25**	4/10	4/10
4	d6			11	d48		
5	d8			12	d250		
6	**d9**	3/6	3/10	13	d113		
7	d511			14	**d3**	5/14	5/10



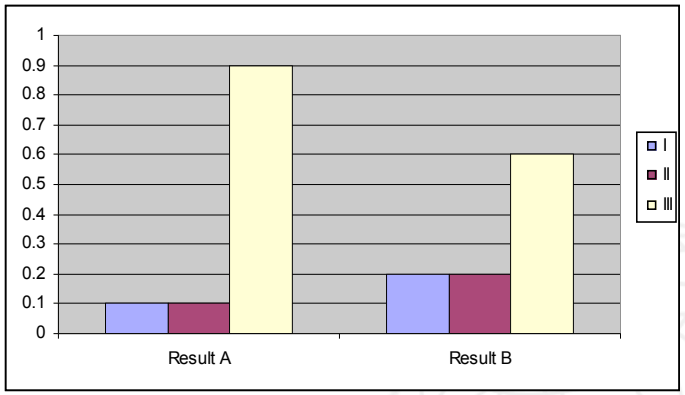
Recall?

- Users „feel“ the precision
- Recall? Not „visible“.
- Even with considerable effort difficult to determine precisely!
  - Number of relevant docs not know.
  - In particular problematic for queries where recall is important (e.g. crime investigations, legal applications, patent search)

Prec at N

- Precision at N (10) documents
  - Clear interpretation
  - Reasonable for web retrieval
  - Little information about the system
  - Position of relevant documents not considered

Which System is better?



2 Systems (A and B), 3 Topics

„IR Psychology“

«The **unhappy customer**, on average, **will tell 27** other people …»

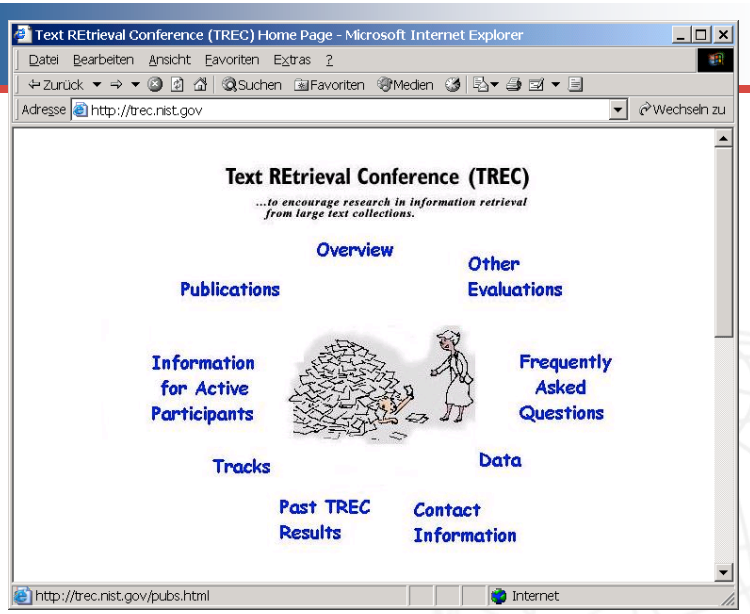
→ ***Bad news travels fast.***

Site search needs to be robust

Avoid bad outliers!

for as many queries as possible

for as many measures as possible



Quelle: TREC homepage http://trec.nist.gov

NTCIR

- Cross-lingual IR asian languages
- Tokio
  - National Institute for Informatics
- Tasks
  - Cross-lingual
    - Chinesisch, Japanisch, Koreanisch -> Englisch
  - Patent-Retrieval
  - Web-Retrieval
  - Question Answering