Information Retrieval Applications

- Application Examples
- Facets of Search
Web Search

Weitere wichtige Vertreter der Frühphase des Information Retrieval waren Mortimer Taube, der das Unitem System entwickelte, Hans Peter Luhn, der das Modell ... Defintion - Geschichte - Retrievalmodelle - Klassifikation von...
do.wikipedia.org/wiki/Information_Retrieval - Im Cache

PDF] 1. Was versteht man unter einem Information Retrieval System?
Dateiformat: PDF/Adobe Acrobat - Schnellansicht
www.ai.wu.ac.at/~wyk/ir/vo/Folien_Kapitel_1.pdf

Abgeschlossene Diplomarbeit: Entwicklung einer ...
Die in einem Hypertext Information Retrieval System (HIRS) verwalteten Objekte sind strukturierte Dokumente, die untereinander wiederum in Beziehung ... www.is.informatik.uni-duisburg.de/dpa/sarr.html - Im Cache - Ähnlich

Information Retrieval Systems - [ Diese Seite übersetzen ]
Information Retrieval Systems ... Virtual Reference Desk · University of Massachusetts Center for Intelligent Information Retrieval · Callan CMU IR Group ...
www.csc.lsu.edu/~kraft/retrieval.html - Im Cache - Ähnlich
**Product Search in Online Shops**

"Icd fernseher 48 zoll"

Verwandte Suchbegriffe: [Icd fernseher 48 zoll](#)

1-16 von 158 Ergebnissen  
Sortieren in [Elektronik & Foto](#)

**Hannspree HA191DPB 48,26cm (19 Zoll) LCD Monitor VGA, DVI (Kontrast dyn. 1000:1), HANNSPREE**

<table>
<thead>
<tr>
<th>EUR</th>
<th>Angebot</th>
</tr>
</thead>
<tbody>
<tr>
<td>109,89</td>
<td>Neu (5 Angebote)</td>
</tr>
<tr>
<td>99,00</td>
<td>Gebraucht (3 Angebote)</td>
</tr>
</tbody>
</table>

*Bestellen Sie in den nächsten 4 Stunden, um den Artikel am Freitag, 26. September zu erhalten.*
*Nur noch 2 Stück auf Lager - jetzt bestellen.*

**Philips 19PFL3606H/12 48 cm (19 Zoll) LCD-Fernseher, Energieeffizienzklasse B (HD-)**

<table>
<thead>
<tr>
<th>EUR</th>
<th>Angebot</th>
</tr>
</thead>
<tbody>
<tr>
<td>225,00</td>
<td>Neu (28 Angebote)</td>
</tr>
<tr>
<td>195,00</td>
<td>Neu (3 Angebote)</td>
</tr>
</tbody>
</table>

*Nur noch 1 Stück auf Lager - jetzt bestellen.*
*Andere Angebote*

**Grundig 46 VLE 8160 SL117 cm (46 Zoll) 3D LED-Backlight-Fernse... Energieeffizienzkl**

<table>
<thead>
<tr>
<th>EUR</th>
<th>Angebot</th>
</tr>
</thead>
<tbody>
<tr>
<td>694,99</td>
<td>Neu (14 Angebote)</td>
</tr>
<tr>
<td>604,84</td>
<td>Gebraucht (2 Angebote)</td>
</tr>
</tbody>
</table>

*Bestellen Sie in den nächsten 7 Stunden, um den Artikel am Freitag, 26. September zu erhalten.*
*Andere Angebote*

**Kostenlose Lieferung möglich**

**Computer & Zubehör: Alle 22 Artikel ansehen**

**Elektronik & Foto: Alle 133 Artikel ansehen**
**Intranet Search**

**Universität Duisburg-Essen**

<table>
<thead>
<tr>
<th>UDE-Suchseite</th>
<th>Google-Web</th>
<th>Google-Bilder</th>
</tr>
</thead>
</table>

**Vorlesung datenbanken** | **default_collection** | **UDE Suche** |

**Suche**

Ergebnisse 1 - 10 von ca. 31 für **Vorlesung datenbanken**. Die Suche dauerte 0.25 Sekunden.

**Weiter**

**Sortieren nach Datum** / **Sortieren nach Relevanz**

---

**Universitätsbibliothek Duisburg-Essen: Schlüsselkompetenzen für ...**
... in nationalen und internationalen **Datenbanken**. Kreditierung: 0,5 ECTS. Studienleistung:
Klausur. Termine: Mi, 09.01.2008, 14:15 - 15:45 Uhr: 1. **Vorlesung** Mi, 16.01 ...  
www.ub.uni-duisburg-essen.de/biblio/schulung/ba.shtml - 45k

**moodle uni-due: Datenverwaltungssysteme und Wissensrepräsentation**
... in der **Vorlesung** werden zunächst die Grundlagen verteilter Systeme ... Architektur 
vertelter Datenbanksysteme; Entwurf verteilter **Datenbanken**; Anfrageverarbeitung; ...  
moodle.uni-duisburg-essen.de/course/category.php?id=75 - 27k

**moodle uni-due: Technik der Rechnernetze**
... **Vorlesung** "Netzmanagement" (2 SWS, 3 CP, WS). ... Verknüpfung mit öffentlicher 
IP-Netzverwaltung / Whois-**Datenbanken**; Pathologische Routingergebnisse. ...  
moodle.uni-duisburg-essen.de/course/category.php?id=76 - 55k

---

**Vorlesungen**

... **Vorlesung** Inhalt. ... **DB (Datenbanken)**, Datenbank - Grundlagen, Anwendungen relationaler 
Datenbanken, Anwendungen objektorientierter **Datenbanken**, Schnittstelle CAD ...
www.uni-due.de/lkb/vorlesungen.shtml - 13k

**Prof. Dr. Rüdiger Schmitt-Beck**

... **Die Vorlesung** gibt einen Überblick der wesentlichen Theorieansätze und ... Kenntnisse 
In Programmierung - vorzugsweise in PHP und mySQL-**Datenbanken** - oder die ...  
www.uni-due.de/politik/schmitt-beck Lehre.php - 76k
Searching in Digital Libraries

1. **Image retrieval: ideas, influences, and trends of the new age**
   - Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang
   - ACM Computing Surveys (CSUR), Volume 40 Issue 2
   - Publisher: ACM
   - Full text available: PDF (2.81 MB)
   - Bibliometrics: Downloads (6 Weeks): 482, Downloads (12 Months): 1801, Citation Count: 2

   We have witnessed great interest and a wealth of promise in content-based image retrieval as an emerging technology. While the last decade laid foundation to such promise, it also paved the way for a large number of new techniques and systems, got many ...

   **Keywords**: Content-based image retrieval, annotation, learning, modeling, tagging

2. **Content-based image retrieval: approaches and trends of the new age**
   - Ritendra Datta, Jia Li, James Z. Wang
   - MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval
   - Publisher: ACM
   - Full text available: PDF (467.64 KB)
   - Bibliometrics: Downloads (6 Weeks): 108, Downloads (12 Months): 728, Citation Count: 14
Multimedia Search

**Morgenstimmung am Neckar** von _darklight_

- 146 Kommentare  ★ 25 Favoriten
- Mit Tag: winter, deutschland, licht, wasser ...
- Aufgenommen in Neckarhausen, Baden-Württemberg (Karte)

**Morgenstimmung** von _muchas641_

- 21 Kommentare  ★ 11 Favoriten
- Mit Tag: korn, neuwirtshaus, nebel, fog ...
- Aufgenommen in Neuwirtshaus, Stuttgart, BW, Deutschland (Karte)

**Morgenstimmung** von _bernd obervossbeck_

- 51 Kommentare  ★ 14 Favoriten
- Mit Tag: morning, autumn, tree, herbst ...

Mehr Fotos von _darklight_ oder sein Profil ansehen.

Mehr Fotos von _muchas641_ oder sein Profil ansehen.

Mehr Fotos von _bernd obervossbeck_ oder sein Profil ansehen.
Example: Cross-lingual search in Google

Nach Websites in anderen Sprachen suchen

Sonnenergie

Seiten durchsuchen, die in folgender Sprache geschrieben sind:

Meine Sprache:

Übersetzte Ergebnisse von englischen Webseiten

Deutsche Übersetzung

Solarpower - Wikipedia, the free encyclopedia
Solar powered electrical generation relies on heat engines and photovoltaics. Solar power's uses are limited only by human ingenuity.

Englischer Originaltext - Englische Ergebnisse ausblenden

Solar power - Wikipedia, the free encyclopedia
Solar power is the generation of electricity from sunlight. This can be direct, as with photovoltaic (PV), or indirect, as with concentrating solar power...

Solar Power
SolarPower.org is dedicated to the rapid deployment of renewable energy and solar power across America. Here you'll find tools, information, and industry...
Facets of Search

Structure

Example: XML retrieval

Our description of the intersystem handoff follows IS-41[2] (GSM follows similar procedures), and we assume network-controlled handoff. Figure 3 illustrates the trunk (voice or data circuit) connection before and after the handoff. A communicating mobile user moves out of the base station served by MSC₁ and enters the area covered by MSC₂. The handoff follows these steps:

- MSC₁ requests MSC₂ to perform handoff measurement. MSC₂ then selects a candidate base station, BS₂, for handoff. That is, MSC₂ finds a base station that covers the mobile phone and has a free radio channel to cover the call. MSC₂ returns the signal-quality parameter values and other information to MSC₁.
- MSC₁ checks if the mobile phone has made too many handoffs or if intersystem handoffs are not available. If so, MSC₁ ends the procedure. Otherwise, MSC₁ asks MSC₂ to set up a voice channel. Suppose that a voice channel is available in BS₂. MSC₂ asks MSC₁ to start the radio link transfer.
- MSC₂ sends the mobile phone a handoff order. The mobile phone tries to synchronize to BS₂. After the mobile phone connects to BS₂, MSC₂ informs MSC₁ that the handoff is successful. MSC₁ then connects the call path (trunk) to MSC₂ and completes the handoff.

Figure 3: Before (a) and after (b) an intersystem handoff.
Example: Similarity search for images

![Image results](https://www.example.com/image)

Options: Standard, Any size, Any time, Gray scale only, Omit same, Reset parameters, Safe search is on
Facets of Search

Objects

Example: People search in 123people
Facets of Search
static/dynamic Content

Example: Twitter search

cdu nrw

Realtime results for cdu nrw
0.04 seconds

1 more results since you started searching. Refresh to see them.

1 minute ago from Echofon · Reply · View Tweet

SteffiLemke: Zehn Wochen vor einer Wahl den Wahlkampfleiter raus schmeißen zu müssen zeugt von inneren Zerfallserscheinungen. #NRW #CDU
2 minutes ago from Echofon · Reply · View Tweet

Mindener: RT @mehr_demokratie: CDU-Generalsekretär Wüst tritt zurück http://bit.ly/c0yDsT (expand) #CDU #NRW #Ruettgers #Wuest
4 minutes ago from web · Reply · View Tweet
Facets of Search

- Language: monolingual, cross-lingual, multilingual
- Structure: atomics, fields, tree structure (e.g. XML), graph (e.g. Web)
- Media: texts, facts, images, audio, video, 3D,…
- Objects: products, people, companies,…
- static/dynamic contents (databases/streams)
Definition of Information Retrieval
Information Retrieval (IR) is about *vagueness* and *uncertainty* in information systems.

**Vagueness:** user cannot give a precise specification of her information need.
Information Retrieval (IR) is about *vagueness* und *uncertainty* in information systems

**Vagueness**: user cannot give a precise specification of her information need

- vague query conditions
Information Retrieval (IR) is about **vagueness** and **uncertainty** in information systems.

**Vagueness:** user cannot give a precise specification of her information need

- vague query conditions
- iterative query formulation
Information Retrieval (IR) is about *vagueness* and *uncertainty* in information systems.

**Vagueness**: user cannot give a precise specification of her information need
- vague query conditions
- iterative query formulation

**Uncertainty**: system has uncertain knowledge about the (content of the) objects in database
Information Retrieval (IR) is about **vagueness** and **uncertainty** in information systems.

**Vagueness:** user cannot give a precise specification of her information need
- vague query conditions
- iterative query formulation

**Uncertainty** system has uncertain knowledge about the (content of the) objects in database
- uncertain representation
  \( \rightarrow \) *wrong answers*
Definition of Information Retrieval

Information Retrieval (IR) is about *vagueness* and *uncertainty* in information systems.

**Vagueness**: user cannot give a precise specification of her information need
- vague query conditions
- iterative query formulation

**Uncertainty**: system has uncertain knowledge about the (content of the) objects in database
- uncertain representation
  - wrong answers
- incomplete representation
  - missing answers
IR = Content-Oriented Search

Narrow Definition of IR

Searching at different abstraction levels:

- **Syntax**  document as sequence of symbols
- **Semantics**  meaning of a text/media object
- **Pragmatics**  usefulness for solving my current problem
Welcome at the Information Engineering Group. Our current work focuses on information retrieval, digital libraries and web-based information systems, with special emphasis on user-oriented research.

**Syntax**  ’digital library’ $\rightsquigarrow$ no match

**Semantics**  ’research area’ $\rightsquigarrow$ match

**Pragmatics**  ’potential project partner for medical information project’?
Retrieval Quality
in contrast to databases, IR system cannot decide if an answer is correct or not

- user has **information need**
- **relevance**: relationship between document and information need
- judged by user
Facets of Relevance

Legend:
- Assessor's/user's cognitive space
- Work task situation
- Cognitive perception of W
- Situational relevance
- Pertinence relevance
- Intellectual topicality
- Algorithmic relevance
- Information need
- Request/query version
- Retrieved information object(s)
- Relevance assessment(s) or interpretation(s)
- Transformation
- IR system
Facets of Relevance

- Situational Relevance: related to the perceived task
- Pertinence relevance: related to the information need
- Intellectual topicality: as judged by human observer
- Algorithmic relevance: system score comparing request/query with object

In the following: Relevance as pertinence/topicality without further distinction
Retrieval metrics

**RET**: set of retrieved documents

**REL**: set of relevant documents in the database

**Precision** $p$: Proportion of relevant among retrieved

**Recall** $r$: Proportion of retrieved among relevant

\[
p = \frac{|REL \cap RET|}{|RET|} \quad r = \frac{|REL \cap RET|}{|REL|}
\]
Retrieval metrics

\( \text{RET} \): set of retrieved documents
\( \text{REL} \): set of relevant documents in the database

Precision \( p \): Proportion of relevant among retrieved
Recall \( r \): Proportion of retrieved among relevant

\[
p = \frac{|\text{REL} \cap \text{RET}|}{|\text{RET}|} \quad r = \frac{|\text{REL} \cap \text{RET}|}{|\text{REL}|}
\]

Example:
20 relevant documents for the current query.
System returns 10 dokumente, of which 8 are relevant.
Retrieval metrics

**RET**: set of retrieved documents

**REL**: set of relevant documents in the database

**Precision** $p$: Proportion of relevant among retrieved

**Recall** $r$: Proportion of retrieved among relevant

\[
p = \frac{|REL \cap RET|}{|RET|} \quad r = \frac{|REL \cap RET|}{|REL|}
\]

**Example:**
20 relevant documents for the current query.
System returns 10 dokumente, of which 8 are relevant.

Precision: $p = \frac{8}{10} = 0.8$
Recall: $r = \frac{8}{20} = 0.4$
Representations

- Semantic Descriptions
- Free Text Search
- Objects, Representations, and Descriptions
Representations

Free text search search in document text
Semantic approach assign semantic descriptions
Semantic Descriptions

classification schemes e.g. hierarchic classification, as in libraries or product catalogs

Tagging users assign tags

Ontologies e.g. OWL: Web Ontology Language
Classification/Ontology Example: DMOZ

DMOZ open directory project

about dmoz | dmoz blog | suggest URL | help | link | editor login

Arts
Movies, Television, Music...

Games
Video Games, RPGs, Gambling...

Kids and Teens
Arts, School Time, Teen Life...

Reference
Maps, Education, Libraries...

Shopping
Clothing, Food, Gifts...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Business
Jobs, Real Estate, Investing...

Health
Fitness, Medicine, Alternative...

Computers
Internet, Software, Hardware...

Home
Family, Consumers, Cooking...

Recreation
Travel, Food, Outdoors, Humor...

Science
Biology, Psychology, Physics...

Sports
Baseball, Soccer, Basketball...
Free Text Search

Problems

Inflection

*computer* – *computers*, *fly* – *flies*
*go* – *goes* – *going*

Derivation

*compute* - *computer* - *computerization* - *computation*

Synonyms

*mobile* – *smartphone*, *table* – *bench* – *board* – *counter*

Polysemes

*bank*, *head*

Compounds

*steamboat*, *testbed*

Phrases

*information retrieval* – *retrieval of information*
Free Text Search
Approaches

- **inflection, derivation** stemming algorithms
  - *computer, computation, computerize* → *comput*

- **synonyms** synonym lexicons

- **compunds** splitting algorithms

- **phrases** adjacency search

Most systems implement only stemming and adjacency search!
Some Inconsistencies and Misidentified Modeling Assumptions in Probabilistic Information Retrieval

WILLIAM S. COOPER
University of California, Berkeley

Research in the probabilistic theory of information retrieval involves the construction of mathematical models based on statistical assumptions. One of the hazards inherent in this kind of theory construction is that the assumptions laid down may be inconsistent in unanticipated ways with the data to which they are applied. Another hazard is that these assumptions may not be chosen on which the derived modeling equations or resulting experiments are actually based. Both kinds of mistakes have been made in past research on probabilistic information retrieval.

One consequence of these errors is that the statistical character of certain probabilistic IR models, including the so-called Binary Independence model, has been seriously misrepresented.

Categories and Subject Descriptors: H.3.1 [Models and Principles]: User/Machine Systems; H.3.3 [Information Storage and Retrieval]: General; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—retrieval models

General Terms: Experimentation, Measurement, Performance, Theory

Additional Key Words and Phrases: Assumptions, bibliographic searching, consistency, document retrieval, independence, logic, modeling

1. INTRODUCTION

Probability theory provides a powerful springboard from which to launch theories of information retrieval and inductive searching. It is, of course, desirable that a formulation be logically powerful. However, such power comes at the price of a certain risk of accidental misuse and abuse. One of the hazards that an IR system designer should be aware of is that of becoming ensnared in statistical simplifying assumptions logically inconsistent with the data from which inferences must be drawn. Another danger is that the fundamental assumptions underlying a theory may be incorrectly stated, and

This is a revised and extended version of a paper presented at the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, October 1991, Chicago, IL.

Author's address: School of Library and Information Studies, University of California, Berkeley, CA 94720. email: w cooper@video.berkeley.edu

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notices are given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

© 1993 ACM 0001-0782/93/0100-0100 $0.00

Research in the probabilistic theory of information retrieval involves the construction of mathematical models. In this kind of theory construction the assumptions laid down ...
Research in the probabilistic theory of information retrieval involves the construction of mathematical models. In this kind of theory construction the assumptions laid down ...

**Stopword removal and stemming:**
research probabil theory informat retriev involv construct mathemat model kind theory construct assume lay down
Example: document text, representation, description

Text:
Research in the probabilistic theory of information retrieval involves the construction of mathematical models. In this kind of theory construction the assumptions laid down ...

Stopword removal and stemming:
research probabil theory informat retriev involv construct mathemat model kind theory construct assume lay down

Representation (Bag of words):
(research,1), (probabil,1), (theory,2), (informat,1), (retriev,1), (involv,1), (construct,2), (mathemat,1), (model,1), (kind,1), (assum,1), (lay,1), (down,1),
Example: document text, representation, description

Text:
Research in the probabilistic theory of information retrieval involves the construction of mathematical models. In this kind of theory construction the assumptions laid down ...

Stopword removal and stemming:
research probabil theory informat retriev involv construct mathemat model kind theory construct assume lay down

Representation (Bag of words):
(research,1), (probabil,1), (theory,2), (informat,1), (retriev,1), (involv,1), (construct,2), (mathemat,1), (model,1), (kind,1), (assum,1), (lay,1), (down,1)

Description:
(research,0.5), (probabil,0.5), (theory,1.0), (informat,0.5), (retriev,0.5), (involv,0.5), (construct,1.0), (mathemat,0.5), (model,0.5), (kind,0.5), (assum,0.5), (lay,0.5), (down,0.5)
Information Retrieval: Introduction

Representations

Objects, Representations, and Descriptions

Conceptual Model

\( q_k \in Q \): query representation
\( q_k^D \in Q^D \): query description
\( d_m \in D \): document representation
\( d_m^D \in D^D \): document description

\( \mathcal{R} \): relevance scale
\( \rho \): retrieval function
Probabilistic Models

- Probabilistic Event Space
- Probability Ranking Principle
- Binary Independence Retrieval Model
- BM25 model
- Learning to Rank
Probabilistic Event space

[Fuhr 92]

\[ D \]

\[ Q \]

\[ q_k \left| q_k \right. \]

\[ d_m \left| d_m \right. \]

\[ Q: \text{Queries} \]
\[ q_k: \text{query} \]
\[ q_k: \text{query rep.} \]

\[ D: \text{Documents} \]
\[ d_m: \text{document} \]
\[ d_m: \text{document rep.} \]
Event space

\[ P(R|q_k, d_m) = 0.5 \]
Event space

Event space: $Q \times D$

single element: query-document pair $(q_k, d_m)$

all elements are equiprobable
Event space

Event space: \( Q \times D \)

single element: query-document pair \((q_k, d_m)\)

all elements are equiprobable

relevance judgement \((q_k, d_m) \in R\)

relevance judgements for different documents w.r.t. the same query are independent of each other
Event space

Event space: $Q \times D$

- single element: query-document pair $(q_k, d_m)$
- all elements are equiprobable

Relevance judgement $(q_k, d_m) \in \mathcal{R}$

- relevance judgements for different documents w.r.t. the same query are independent of each other

Probability of relevance $P(\text{rel}|q_k, d_m)$:

- probability of an element of $(q_k, d_m)$ being relevant

  - regard collections as samples of possibly infinite sets
  - poor representation of retrieval objects: single representation may stand for a number of different objects.
defines optimum retrieval for probabilistic models:
rank documents according to decreasing values of the

probability of relevance \( P(\text{rel}|q, d) \)
defines optimum retrieval for probabilistic models:
rank documents according to decreasing values of the probability of relevance $P(\text{rel}|q, d)$

**Advantage:**
PRP yields
- optimum retrieval quality
- minimum retrieval costs
Binary Independence Retrieval Model

represent queries and documents as sets of terms
\( T = \{ t_1, \ldots, t_n \} \) set of terms in the collection

\[ q \in Q : \text{query representation} \]
\[ d_m \in D : \text{document representation} \]

\( q^T \): set of query terms
\( d_m^T \): set of document terms
represent queries and documents as sets of terms
\( T = \{t_1, \ldots, t_n\} \) set of terms in the collection

\[ q \in Q: \text{query representation} \quad q^T: \text{set of query terms} \]

\[ d_m \in D: \text{document representation} \quad d_m^T: \text{set of document terms} \]

simple retrieval function: Coordination level match

\[ \varrho_{COORD}(q, d_m) = |q^T \cap d_m^T| \]
Binary Independence Retrieval Model

represent queries and documents as sets of terms
\( T = \{ t_1, \ldots, t_n \} \) set of terms in the collection

\[ q, q^T, d, d^T \]

simple retrieval function: Coordination level match
\[ \varrho_{COORD}(q, d_m) = |q^T \cap d^T_m| \]

Binary independence retrieval (BIR) model: assign weights to query terms
\[ \varrho_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d^T_m} c_i \]
Binary Independence Retrieval Model

\[ \mathcal{Q}_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_i, \quad c_i = \log \frac{p_i(1 - s_i)}{s_i(1 - p_i)} \]
Information Retrieval: Introduction
Probabilistic Models
Binary Independence Retrieval Model

\[ \varrho_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_i, \quad c_i = \log \frac{p_i(1 - s_i)}{s_i(1 - p_i)} \]

\( p_i = P(t_i|\text{rel}) \): prob. that \( t_i \) occurs in arbitrary relevant doc.
\( s_i = P(t_i|\neg\text{rel}) \): prob. that \( t_i \) occurs in arbitrary nonrelevant doc.
Parameter estimation
Relevance Feedback

<table>
<thead>
<tr>
<th></th>
<th>$t_i$ occurs</th>
<th>$\neg t_i$ occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevant</td>
<td>$r_i$</td>
<td>$R - r_i$</td>
</tr>
<tr>
<td>$\neg$ relevant</td>
<td>$n_i - r_i$</td>
<td>$N - n_i - R + r_i$</td>
</tr>
<tr>
<td></td>
<td>$n_i$</td>
<td>$N - n_i$</td>
</tr>
</tbody>
</table>

$p_i = P(t_i|rel)$  prob. that $t_i$ occurs in arbitrary relevant doc.

$$p_i \approx \frac{r_i}{R}$$

$s_i = P(t_i|\neg rel)$  prob. that $t_i$ occurs in arbitrary nonrelevant doc.

$$s_i \approx \frac{n_i - r_i}{N - R} \approx \frac{n_i}{N}$$
Parameter estimation w/o relevance feedback

\( N \) - \# documents in the collection
\( n_i \) - \# documents containing term \( t_i \)

\( s_i = P(t_i|\bar{rel}) \) prob. that \( t_i \) occurs in arbitrary nonrelevant doc.

\[ s_i \approx \frac{n_i}{N} \]

\( p_i = P(t_i|rel) \) prob. that \( t_i \) occurs in arbitrary relevant doc.
assume constant value: \( p = 0.5 \)
Parameter estimation w/o relevance feedback

\( N \) - \# documents in the collection

\( n_i \) - \# documents containing term \( t_i \)

\( s_i = P(t_i|\neg \text{rel}) \) prob. that \( t_i \) occurs in arbitrary nonrelevant doc..

\[ s_i \approx \frac{n_i}{N} \]

\( p_i = P(t_i|\text{rel}) \) prob. that \( t_i \) occurs in arbitrary relevant doc.

assume constant value: \( p = 0.5 \)

\[ c_i = \log \frac{p_i(1 - s_i)}{s_i(1 - p_i)} = \log \frac{p}{1 - p} + \log \frac{1 - s_i}{s_i} \]

\[ = 0 + \log \frac{N - n_i}{n_i} \approx \log \frac{N}{n_i} \]

IDF (inverse document frequency) weight: \( \log \frac{N}{n_i} \)
BM25

[Robertson et al 95]
heuristic extension of the BIR model
from binary to weighted indexing
(consideration of within-document frequency $tf$)
tf*idf Weighting

- originally developed for (non-probabilistic) vector space model
- set of heuristics: the weight of a term should be higher...
  1. the less frequent the term occurs in the collection (inverse document frequency, \( idf \) — see above)
  2. the more often the term occurs in the document (\( tf \))
  3. the shorter the document
From binary to weighted Indexing

- $l_m$: document length (# tokens in $d_m$)
- $al$: average document length in $D$
- $tf_{mi}$: occurrence frequency of $t_i$ in $d_m$.
- $b$: weight of length normalization, $0 \leq b \leq 1$
- $k$: weight of occurrence frequency

length normalization: $B = \left(1 - b\right) + b \frac{l_m}{al}$

normalized within-document frequency: $ntf_{mi} = \frac{tf_{mi}}{B}$

BM25 weight: $u_{mi} = \frac{ntf_{mi}}{k + ntf_{mi}}$

$= \frac{tf_{mi}}{k \left(1 - b\right) + b \frac{l_m}{al}} + tf_{mi}$
Parameter learning in IR

[Fuhr 92]

Learning approaches in IR

query-related learning

document-related learning

feature-related learning
Learning to Rank for Web Searches

Page rank
Query and doc. features: BM25, term locations, word distance, ...
Features of query and anchor text
Information about the user and his social network

Machine Learning/ classification methods

\[ P(R|\bar{x}(q,d)) \]
Interactive Retrieval

- Search models
- Anomalous State of Knowledge
- Ingwersen’s Cognitive Model
Search models
Classical search process model

1. Information Need
2. Query
3. Send to System
4. Receive Results
5. Evaluate Results
6. Reformulate
7. Done?
   a. Yes: Stop
   b. No: Go back to 1.
Empirical studies

- Information search consists of a sequence of connected, but different searches.
- Search result may trigger new searches.
- Only task context remains the same.
- Main goal of a search is accumulated learning and collection of new information while searching.
Search models
Berry picking-Model

[Bates 90]
- continuous change of information need and queries during search
- information need cannot be satisfied by a single result set
- instead: sequence of selections and collection of pieces of information during search
Anomalous State of Knowledge (ASK)(1)

[Belkin 80]

**Classic IR systems: ”best match” principle**

- system returns those documents that fit best to the representation of the information need (e.g. query statement)
- only feasible, if user can give precise specification of her information need (like e.g. in DBMS)
Anomalous State of Knowledge (ASK)(2)

ASK-Hypothesis

information need results from user's anomalous state of knowledge (ASK)
user is unable to precisely specify information need for removing the ASK
instead: describe ASK requires capture of cognitive and situation-specific aspects for resolving this anomaly
Anomalous State of Knowledge (ASK) (2)

ASK-Hypothesis

- information need results from user’s anomalous state of knowledge (ASK)
Anomalous State of Knowledge (ASK) (2)

ASK-Hypothesis

- information need results from user’s *anomalous state of knowledge (ASK)*
- user is unable to precisely specify information need for removing the ASK
Anomalous State of Knowledge (ASK)(2)

**ASK-Hypothesis**

- information need results from user’s *anomalous state of knowledge (ASK)*
- user is unable to precisely specify information need for removing the ASK
- instead: describe ASK
Anomalous State of Knowledge (ASK)(2)

**ASK-Hypothesis**

- information need results from user’s *anomalous state of knowledge (ASK)*
- user is unable to precisely specify information need for removing the ASK
- instead: describe ASK
- requires capture of cognitive and situation-specific aspects for resolving this anomaly
Ingwersen’s Cognitive Model

Information objects

IT: Engines
Logics
Algorithms

Interface

Cognitive Actor(s)
(team)

Organiz.

Social Context

Cultural

Cognitive transformations and influence

Interactive communications of cognitive structures