Information Retrieval: Introduction

Norbert Fuhr

May 15, 2015
Information Retrieval Applications

- Application Examples
- Facets of Search
Web Search

Weitere wichtige Vertreter der Frühphase des Information Retrieval waren Mortimer Taube, der das Unilever-System entwickelte, Hans Peter Luhn, der das Modell ...  
Definition - Geschichte - Retrievalmodelle - Klassifikation von ...
de.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/fofolien_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

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

EUR 109,89
Bestellen Sie in den nächsten 4 Stunden, um den Artikel am Freitag, 28. September zu erhalten.
Nur noch 2 Stück auf Lager - jetzt bestellen.

Andere Angebote
EUR 99,00 neu (56 Angebote)
EUR 96,34 gebraucht (3 Angebote)

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

EUR 339,99 EUR 225,00
Nur noch 1 Stück auf Lager - jetzt bestellen.

Andere Angebote
EUR 195,00 neu (3 Angebote)

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

EUR 1,099,90 EUR 694,99
Bestellen Sie in den nächsten 7 Stunden, um den Artikel am Freitag, 28. September zu erhalten.

Andere Angebote
EUR 694,99 neu (14 Angebote)
EUR 694,64 gebraucht (2 Angebote)
Intranet Search

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 - 46k

moodle uni-due: Datenverwaltungssysteme und Wissensrepräsentation
... In der Vorlesung werden zunächst die Grundlagen verteilter Systeme ... Architektur
verteilter 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 / Whols-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/ikb/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
1. **Image retrieval: ideas, influences, and trends of the new age**
   
   Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang
   
   April 2008, 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
   
   **Keywords:** Content-based image retrieval, annotation, learning, modeling, tagging

   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 ...

2. **Content-based image retrieval: approaches and trends of the new age**
   
   Ritendra Datta, Jia Li, James Z. Wang
   
   November 2005, 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

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

Morgenstimmung am Neckar von _darklight

Morgenstimmung von muchas641

Morgenstimmung von bernd obervosbeck
Facets of Search
Language

Example: Cross-lingual search in Google

Nach Websites in anderen Sprachen suchen

Sonnenergie

Seiten durchsuchen, die in folgender Sprache geschrieben sind:
Deutsche Sprache:

Übersetzte Ergebnisse von englischen Webseiten

Deutsche Übersetzung

Solar energy - Wikipedia, the free encyclopedia
Solar powered electrical generation relies on heat engines and photovoltaics. Solar energy'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 photovoltaics (PV), or indirect as with concentrating 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 ...

www.solarpower.org/- 27k - Cached
Example: XML retrieval

**Figure 3:** Before (a) and after (b) an inter-system handoff.
Facets of Search

Media

Example: Similarity search for images
Facets of Search

Example: People search in 123people

Norbert Fuhrs Fotos (26)

Email Adressen (9)

1 2 weiter>>

Telefonbuch (11)

1 2 weiter>>
Facets of Search
static/dynamic Content

Example: Twitter search

cdu nrw

Realtime results for cdu nrw

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

- 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)
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)
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)
Information Retrieval (IR) is about \textit{vagueness} and \textit{uncertainty} in information systems.

\textbf{Vagueness:} user cannot give a precise specification of her information need
- vague query conditions
- iterative query formulation

\textbf{Uncertainty} system has uncertain knowledge about the (content of the) objects in database
- uncertain representation \(\leadsto\) wrong answers
- incomplete representation \(\leadsto\) missing answers
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
- iterative query formulation

**Uncertainty** system has uncertain knowledge about the (content of the) objects in database
- uncertain representation
  - \( \rightsquigarrow \) wrong answers
- incomplete representation
  - \( \rightsquigarrow \) missing answers
Information Retrieval (IR) is about \textit{vagueness} und \textit{uncertainty} in information systems.

\textbf{Vagueness:} user cannot give a precise specification of her information need
- vague query conditions
- iterative query formulation

\textbf{Uncertainty} system has uncertain knowledge about the (content of the) objects in database
- uncertain representation \(\leadsto\text{ wrong answers}\)
- incomplete representation \(\leadsto\text{ missing answers}\)
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
Syntax, Semantics and Pragmatics in Image Search

**Syntax**  colours, contours, textures

**Semantics**  objects, properties, relationships

**Pragmatics**  task-related
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’ \(\sim\) no match
Semantics  'research area’ \(\sim\) match
Pragmatics  'potential project partner for medical information project’?
Retrieval Quality
Retrieval Quality
The concept of relevance

- 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:
- Circle: Assessor’s/user’s cognitive space
- W: Work task situation
- CW: Cognitive perception of W
- SR: Situational relevance
- P: Pertinence relevance
- IT: Intellectual topicality
- A: Algorithmic relevance
- N: Information need
- r/q: request/query version
- O: Retrieved information object(s)
- O-Or: Collection of objects
- Collection: 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
Retrieval metrics

**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: $r = \frac{8}{20} = 0.4$
Retrieval metrics

**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: \( r = \frac{8}{20} = 0.4 \)
Retrieval metrics

**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
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

**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...

**Business**
Jobs, Real Estate, Investing...

**Health**
Fitness, Medicine, Alternative...

**Computers**
Internet, Software, Hardware...

**Home**
Family, Consumers, Cooking...

**Recreation**
Travel, Food, Outdoors, Humor...

**Regional**
US, Canada, UK, Europe...

**Science**
Biology, Psychology, Physics...

**Sports**
Baseball, Soccer, Basketball...

**World**
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...
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
- compounds 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 the stated assumptions may not be clear 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.1.2 [Models and Principles]: User/Machine Systems; H.3.0 [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 reasoning. 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: wcoop@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.
© 1995 ACM 0003-0003/95/0100-0103-1110

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)
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)
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)
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)
Conceptual Model

\[ \begin{align*}
Q & \xrightarrow{\alpha_Q} Q & \xrightarrow{\beta_Q} Q^D \\
D & \xrightarrow{\alpha_D} D & \xrightarrow{\beta_D} D^D
\end{align*} \]

- \( q_k \in Q \): query representation
- \( q_k \in Q \): query
- \( q_k^D \in Q^D \): query description
- \( d_m \in D \): document representation
- \( d_m \in D \): document
- \( d_m^D \in D^D \): document description
- \( \mathcal{R} \): relevance scale
- \( \varrho \): retrieval function
Probabilistic Models

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

[Fuhr 92]

\[
\begin{array}{c|c|c}
\hline
D & q_k & q_k \\
\hline
Q & \hline
\hline
\hline
d_m & \hline
\hline
\hline
D: & \text{Documents} & \\
\hline
Q: & \text{Queries} & \\
q_k & \text{query} & \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c}
\hline
D & q_k & q_k \\
\hline
Q & \hline
\hline
\hline
d_m & \hline
\hline
\hline
D: & \text{Documents} & \\
\hline
Q: & \text{Queries} & \\
q_k & \text{query rep.} & \\
\hline
\end{array}
\]
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

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 a 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.
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 a 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.
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 $P(\text{rel} \mid q_k, d_m)$:
- 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 $P(\text{rel}|q, d)$

Advantage:

PRP yields

- optimum retrieval quality
- minimum retrieval costs
Probability Ranking Principle

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
PRP Example

System computes the following probabilities of relevance $P(R|q, d)$:
(0.9, 0.8, 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.0)
User regards the first three documents only

1. What is the expected precision?
2. What is the expected recall?

1. $p = (0.9 + 0.8 + 0.5)/3 = 0.73$
2. $\sum_i P(R|q, d_i) = 4, \quad r = (0.9 + 0.8 + 0.5)/4 = 0.55$
PRP Example

System computes the following probabilities of relevance $P(R|q, d)$:
$(0.9, 0.8, 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.0)$
User regards the first three documents only

1. What is the expected precision?
   - $p = (0.9 + 0.8 + 0.5)/3 = 0.73$

2. What is the expected recall?
   - $r = (0.9 + 0.8 + 0.5)/4 = 0.55$
System computes the following probabilities of relevance $P(R|q, d)$:
(0.9, 0.8, 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.0)
User regards the first three documents only

1. What is the expected precision?
   $p = (0.9 + 0.8 + 0.5)/3 = 0.73$

2. What is the expected recall?
   $\sum_i P(R|q, d_i) = 4, \quad r = (0.9 + 0.8 + 0.5)/4 = 0.55$
represent queries and documents as sets of terms
\( T = \{ t_1, \ldots, t_n \} \) set of terms in the collection

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

simple retrieval function: Coordination level match

\[ \varrho_{\text{COORD}}(q, d_m) = |q^T \cap d_m^T| \]

Binary independence retrieval (BIR) model:
assign weights to query terms

\[ \varrho_{\text{BIR}}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_{i} \]
Binary Independence Retrieval Model

represent queries and documents as sets of terms

\[ T = \{ t_1, \ldots, t_n \} \] set of terms in the collection

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

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

simple retrieval function: **Coordination level match**

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

Binary independence retrieval (BIR) model:
assign weights to query terms

\[ \varrho_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_i \]
Binary Independence Retrieval Model

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

\( q \in Q: \) query representation
\( d_m \in D: \) document representation

simple retrieval function: Coordination level match

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

Binary independence retrieval (BIR) model:
assign weights to query terms

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

$p_i = P(t_i|\text{rel})$: prob. that $t_i$ occurs in arbitrary relevant doc.
$s_i = P(t_i|\overline{\text{rel}})$: prob. that $t_i$ occurs in arbitrary nonrelevant doc.
\[ \rho_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d_m^T} c_i, \quad c_i = \log \frac{p_i(1 - s_i)}{s_i(1 - p_i)} \]

- \( p_i = P(t_i|rel) \): prob. that \( t_i \) occurs in arbitrary relevant doc.
- \( s_i = P(t_i|\bar{rel}) \): 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>$n_i$</td>
<td>$N - n_i$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

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

\[ p_i \approx \frac{r_i}{R} \]

$s_i = P(t_i|\bar{\text{rel}})$  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: \( p = 0.5 \)

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

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

IDF (inverse document frequency) weight: \( \log \frac{N}{n_i} \)
Parameter estimation w/o relevance feedback

$N$ - # documents in the collection

$n_i$ - # documents containing term $t_i$

$s_i = P(t_i|\neg rel)$ 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: $p = 0.5$

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

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

IDF (inverse document frequency) weight: $\log \frac{N}{n_i}$
[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 (number of 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: $B = \left( (1 - b) + b \frac{l_m}{al} \right)$

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

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

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

Click-through data

\[ P(R|\tilde{x}(q,d)) \]
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

- Reformulate
  - No
    - Done?
      - Yes
        - Stop
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.
Information Retrieval: Introduction
Interactive Retrieval
Anomalous State of Knowledge

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
Information Retrieval: Introduction
Interactive Retrieval
Anomalous State of Knowledge

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)*
- 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)*
- 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