Information Retrieval: Introduction

Norbert Fuhr

April 28, 2014

Information Retrieval Applications

Application Examples
Facets of Search

Web Search

Product Search in Online Shops
Intranet Search

Searching in Digital Libraries

Multimedia Search

Facets of Search

Language

Example: Cross-lingual search in Google

Deutsche Übersetzung

Solar Power: Siehe die folgende Grafik.

Solar Power: Siehe die folgende Grafik.

Solar Power: Siehe die folgende Grafik.
Facets of Searching

Structure
Example: XML retrieval

Facets of Searching

Media
Example: Similarity search for images

Facets of Searching

Objects
Example: People search in 123people

Facets of Searching

Static/dynamic Content
Example: Twitter search
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 ($\Rightarrow$ wrong answers)
- incomplete representation ($\Rightarrow$ 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' ⇦ no match
Semantics: 'research area' ⇦ match
Pragmatics: 'potential project partner for medical information project'?

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
Facets of Relevance

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

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

Retrieval metrics

\[
\begin{align*}
\text{RET: } & \text{ set of retrieved documents} \\
\text{REL: } & \text{ set of relevant documents in the database} \\
\text{Precision } p: & \text{ Proportion of relevant among retrieved} \\
\text{Recall } r: & \text{ Proportion of retrieved among relevant} \\
\end{align*}
\]

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

Example:
20 relevant documents for the current query. System returns 10 documents, of which 8 are relevant.
Precision: \( p = \frac{8}{10} = 0.8 \)
Recall: \( r = \frac{8}{20} = 0.4 \)

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

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

Classification/Ontology Example: DMOZ

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!
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

Probabilistic Models

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

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.

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
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
- \( q^T \): set of query terms
- \( d_m \in D \): document representation
- \( d^T_m \): set of document terms

simple retrieval function: Coordination level match

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

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

\[ \rho_{BIR}(q, d_m) = \sum_{t_i \in q^T \cap d^T_m} c_i \]

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 )</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>

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

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

\[ s_i \approx \frac{n_i - r_i}{N - R} \approx \frac{n_i}{N} \]

\[ r_i \]

- \( N \) - # documents in the collection
- \( n_i \) - # documents containing term \( t_i \)
- \( p_i = P(t_i|\text{rel}) \): prob. that \( t_i \) occurs in arbitrary relevant doc.
- \( s_i = P(t_i|\bar{\text{rel}}) \): prob. that \( t_i \) occurs in arbitrary nonrelevant doc.

IDF (inverse document frequency) weight: \( \log N/n_i \)
BM25

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

\[ u_{mi} = \frac{ntf_{mi}}{k + ntf_{mi}} \]
\[ = \frac{tf_{mi}}{k ((1 - b) + b \frac{l_m}{al}) + tf_{mi}} \]

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, \( 0 < b \leq 1 \)
- \( k \): weight of occurrence frequency

\[ B = (1 - b) + b \frac{l_m}{al} \]

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

Parameter learning in IR

[Fuhr 92]
Learning to Rank for Web Searches

Interactive Retrieval

Search models
- Anomalous State of Knowledge
- Ingwersen’s Cognitive Model

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
Classical search process model
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

Ingwersen’s Cognitive Model