IV. Retrieval Models

- Overview of Retrieval Models
- Empirical Models
- Boolean Retrieval
- Vector Space Model
- Probabilistic Models
- Binary Independence Model
- Okapi BM25
- Hidden Variable Models
- Latent Semantic Indexing
- Explicit Semantic Analysis
- Generative Models
- Language Models
- Divergence From Randomness
- Combining Evidence
- Web Search
- Learning to Rank
Overview of Retrieval Models

Document Views

Information retrieval requires modeling and representing documents on a computer. We distinguish three orthogonal views on a document’s content:

1. **Layout view**
   Presentation of a document on a (two-dimensional) medium.

2. **Structural / logical view**
   Composition and logical structure of a document. Example:
   \begin{verbatim}
   \documentclass[twocolumn,german]{article}
   \title{...}
   \author{...}
   \section{...}
   \end{verbatim}

3. **Semantic view**
   The meaning of a document or its message, allowing for its interpretation.
Overview of Retrieval Models

Retrieval Models

- Conceptual model
  - Layout view
  - Structural / logical view
  - Semantic view

Formalized query

Real document: $d \in D$

Information need

Query: $q \in Q$
Overview of Retrieval Models

Retrieval Models

- Conceptual model
  - Layout view
  - Structural / logical view
  - Semantic view

Computer representation of the document: \( d \in D \)

- Computer representation of the query: \( q \in Q \)

- Real document: \( d \in D \)
- Information need
- Query

IR:IV-4 Retrieval Models © HAGEN/POTTHAST/STEIN 2014-2020
Overview of Retrieval Models

Retrieval Models

Conceptual model
- Layout view
- Structural / logical view
- Semantic view

Computer representation of the document
\( d \in D \)

Relevance scoring
\( \rho(q,d) \)

Linguistic Theory

Real document
\( d \in D \)

Information need

Query
\( q \in Q \)

Computer representation of the query

IR:IV-5 Retrieval Models © HAGEN/POTTHAST/STEIN 2014-2020
Overview of Retrieval Models

Retrieval Models

- Conceptual model
  - Layout view
  - Structural / logical view
  - Semantic view

- Computer representation of the document
  - Real document
  - Information need

- Relevance scoring
  - Linguistic Theory

- Retrieval model \( \mathcal{R} \)
  - Query
  - Computer representation of the query
  - \( d \in D \)
  - \( q \in Q \)
  - \( \rho(q,d) \)

IR:IV-6 Retrieval Models © HAGEN/POTTHAST/STEIN 2014-2020
Overview of Retrieval Models

**Definition 1 (Retrieval Model, Relevance Function)**

Let $D$ denote the set of documents and $Q$ the set of queries. A retrieval model $\mathcal{R}$ for $D, Q$ is a tuple $\langle D, Q, \rho \rangle$ defined as follows:

1. $D$ is the set of document representations, where $d \in D$ represents $d \in D$. It may encode information from the layout view, the logical view, and the semantic view.

2. $Q$ is the set of query representations.
Overview of Retrieval Models

Definition 1 (Retrieval Model, Relevance Function)
Let \( D \) denote the set of documents and \( Q \) the set of queries. A retrieval model \( \mathcal{R} \) for \( D, Q \) is a tuple \( \langle D, Q, \rho \rangle \) defined as follows:

1. \( D \) is the set of document representations, where \( d \in D \) represents \( d \in D \).
   It may encode information from the layout view, the logical view, and the semantic view.

2. \( Q \) is the set of query representations.

3. \( \rho(q, d) \) denotes a relevance function, which quantifies the relevance between a query \( q \) and a document \( d \) via their representations \( q \in Q \) and \( d \in D \):
   \[
   \rho : Q \times D \rightarrow R
   \]
   The values computed by \( \rho \) are called relevance scores.

\( \mathcal{R} \) formalizes a certain principle, paradigm, or linguistic theory of retrieval.
Remarks:

- A document representation encompasses certain elements and specific aspects of a real document. Examples for document representations include feature vectors, feature trees, and fingerprints.

- A retrieval model provides the theoretical foundations of how human information needs can be satisfied by drawing information from the three views. Examples for retrieval models include the vector space model, the binary independence model, and latent semantic indexing.

- An alternative name for a retrieval model is retrieval strategy.

- Most retrieval models are based on the semantic view of documents.

- An intensional definition of the sets $Q$ and $D$ can be given as functions $\alpha_Q : Q \rightarrow Q$ and $\alpha_D : D \rightarrow D$. [Fuhr 2004]
Overview of Retrieval Models

History of Retrieval Models  [Stein 2013]

Empirical Models
- Boolean
- VSM
- FuzzySet
- GVSM
- LSI
- Genre
- SuffixTree
- DivRand
- WebGenre
- CL-ESA
- ESA

Probabilistic Models
- ProbabilityIndex
- 2-Poisson
- BIM
- BII
- Inquery
- BeliefNet
- BestMatch
- LanguageModel
- LDA
- MixitureUnigram
- pLSI
- Doc2Vec

Language Models

IR:IV-10  Retrieval Models

©HAGEN/POTTHAST/STEIN 2014-2020
Overview of Retrieval Models

Document Modeling

Deconstruction models

Reconstruction models