Title: Retrieval Models

Name: Benno Stein, Tim Gollub, Maik Anderka

Affil./Addr.: Bauhaus-Universität Weimar

99421 Weimar, Germany

<first>.<last>@uni-weimar.de

Retrieval Models

Synonyms

Document model, Document representation, Document indexing

Glossary

Feature: A characteristic property of a document. Usually, a document's terms are used as features, but virtually every measurable document property can be chosen, such as word classes, average sentence lengths, principal components of term-document-occurrence matrices, term synonyms, etc.

Information need: Specifically here: A lack of information or knowledge that can be satisfied by a text document.

Query: Specifically here: A small set of words that expresses a user's information need.

Relevance: The extent to which a document is capable to satisfy an information need. Within probabilistic retrieval models, relevance is modeled as a binary random variable.

Definition

Retrieval *models* provide the formal means to address (information) retrieval *tasks* with the aid of a computer. A retrieval task is given if an information need is to be satisfied

against an information source. More specifically, the information need is represented as a term query provided by a user, the information source is given in form of a text document collection, and the solution of the retrieval task is a subset of such documents of the collection, which the user considers as relevant with respect to the query. Though a broad range of retrieval tasks can be imagined, including all kinds of multimedia queries and multimedia collections (consider for example "query by humming" or medical image retrieval), the term "retrieval model" is predominantly used in the aforementioned narrow sense. Retrieval models in this sense are based on a linguistic theory and can be considered as heuristics that operationalize the *probability ranking principle* (Robertson, 1997): "Given a query q, the ranking of documents according to their probabilities of being relevant to q leads to the optimum retrieval performance." The principle cannot be applied to all kinds of retrieval tasks. In comment ranking, for example, the differential information gain must be considered.

Retrieval models can be classified according to the linguistic theory they are based upon. In the literature a distinction between *empirical models*, *probabilistic models*, and *language models* is often made, which is rooted in the query-oriented understanding of retrieval tasks but which also has historical reasons.

1. Empirical models, sometimes referred to as vector space models, focus on the document representation (Salton and McGill, 1983). Both documents and queries are considered as high-dimensional vectors in the Euclidean space, whereas a compatible representation is presumed: a particular document term or query term is always associated with the same dimension, whereas the term importance is specified by a weight. Usually, the cosine of the angle between two such vectors is used to quantify their similarity; in particular, the concept of similarity is put on a level with the concept of relevance. Empirical models can be distinguished with regard to the di-

mensions that are considered (features that are chosen) and how these dimensions (features) are weighted.

2. Probabilistic models strive for an explicit modeling of the concept of relevance. Statistics comes into play in order to estimate the probability of the event that a document is relevant for a given information need. Most probabilistic models employ conditional probabilities to quantify document relevance under term occurrence.
3. Language models are based on the idea of language generation as it is used in speech recognition systems. A language-based retrieval model is computed specifically for each document in a collection and is usually term-based. Given a query q, document ranking happens according to the generation probability of q under the language model of the respective document.

Historical Background

Figure 1 illustrates the historical development of well-known retrieval models. From each of the three modeling paradigms (empirical models, probabilistic models, language models) selected representatives are in the following characterized along with the respective publications.

The Boolean retrieval model uses binary term weights, and a query is a Boolean expression with terms as operands. Drawbacks of the Boolean model include its simplistic weighting scheme, its restriction to exact matches, and that no document ranking is possible. The Vector Space Model (VSM) and its variants consider documents and queries embedded in the Euclidean space (see above). Key problem of these kinds of models is the term weighting. Salton et al (1975) proposed the $tf \cdot idf$-scheme, which combines the term frequency tf (the number of term occurrences in a document) with the inverse document frequency idf (the inverse of the number of documents that contain this term). The Latent Semantic Indexing (LSI) model was developed to improve

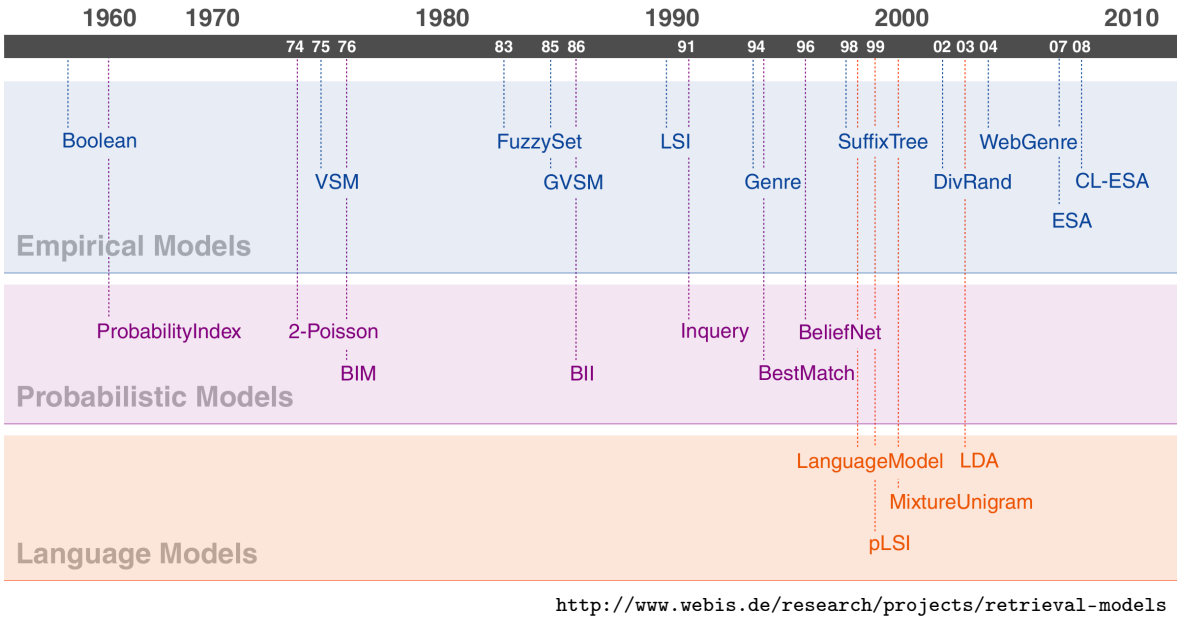

Fig. 1. Historical development of retrieval models, organized according to three paradigms: empirical models, probabilistic models, and language models.

query interpretation and semantic-based matching (Deerwester et al, 1990). E.g., a document d should match a query even if the user specified valid synonyms that do not occur in d. The LSI model attempts to achieve such effects by projecting documents and queries in a "semantic space", which is constructed by a singular value decomposition of the term-document-matrix. The Explicit Semantic Analysis (ESA) model was introduced to compute the semantic relatedness of natural language texts (Gabrilovich and Markovitch, 2007). The model represents a document d as a high-dimensional vector whose the dimensions quantify the pairwise similarities between d and the documents of some reference collection such as Wikipedia. Potthast et al (2008) demonstrated how the ESA principles are applied to develop a very effective cross-language retrieval approach, the so-called CL-ESA model.

Under the Binary Independence Model (BIM) the documents are ranked by decreasing probability of relevance (Robertson and Sparck-Jones, 1976). The model is based on two assumptions which allow for a practical estimation of the required probabilities: documents and queries are represented under a Boolean model, and, the

terms are modeled as occurring independently of each other. The Best Match (BM) model computes the relevance of a document to a query based on the frequencies of the query terms appearing in the document and their inverse document frequencies (Robertson and Walker, 1994). Three parameters tune the influence of the document length, the document term frequency, and the query term frequency in the model. The Best Match model belongs to the most effective retrieval models in the Text Retrieval Conference (TREC) series.

The Language Modeling approach to information retrieval was proposed by Ponte and Croft (1998); the idea is to rank documents by the generation probabilities for a given query (see above). The algorithmic core of the model is a maximum likelihood estimation of the probability of a query term under a document's term distribution. The Latent Dirichlet Allocation (LDA) model is a sophisticated generative model in the context of probabilistic topic modeling. Under this model it is assumed that documents are composed as a mixture of latent topics, where each topic is specified as a probability distribution over words. The mixture is generated by sampling from a Dirichlet distribution.

Cross-References

00382 Analysis and Mining of Tags, (Micro-)Blogs, and Virtual Communities

00056 Data Mining

00141 Distance and Similarity Measures

00352 Document Topic Identification

00142 Eigenvalues, Singular Value Decomposition

00353 Microtext Processing

00034 Mining Trends in the Blogosphere

00188 Social Media Mining and Knowledge Discovery

00261 Social Web Search

00170 Theory of Probability, Basics and Fundamental

00171 Theory of Statistics, Basics and Fundamental

00279 User Reviews Ranking

References

Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R (1990) Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6):391–407

Gabrilovich E, Markovitch S (2007) Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In: Veloso MM (ed) IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, pp 1606–1611

Ponte J, Croft W (1998) A language modeling approach to information retrieval. In: SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York, NY, USA, pp 275–281,

Potthast M, Stein B, Anderka M (2008) A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald C, Ounis I, Plachouras V, Ruthven I, White R (eds) Advances in Information Retrieval. 30th European Conference on IR Research (ECIR 08), Springer, Berlin Heidelberg New York, Lecture Notes in Computer Science, vol 4956, pp 522–530,

Robertson S (1997) The probability ranking principle in IR. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Robertson S, Sparck-Jones K (1976) Relevance Weighting of Search Terms. American Society for Information Science 27(3):129–146

Robertson S, Walker S (1994) Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., New York, NY, USA, pp 232–241

Salton G, McGill M (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York

Salton G, Wong A, Yang C (1975) A Vector Space Model for Automatic Indexing. Commun ACM 18(11):613–620