Retrieval Models

Benno Stein, Tim Gollub, and Maik Anderka

1 Synonyms

Document models, Document representations, Relevance Functions

2 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 set of text documents.

Query Specifically here: A small set of terms 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.

3 Definition

A retrieval *model* provides a formal means to address (information) retrieval *tasks* with the aid of a computer.

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

4 Introduction

A retrieval task is given if an information need is to be satisfied with an information resource. More specifically, the information need is represented as a term query provided by a user, the information resource 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.

5 Key Points

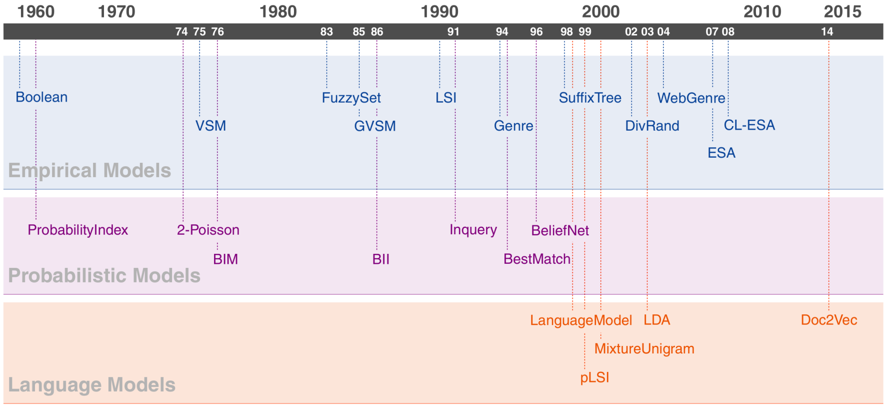
Retrieval models can be classified according to the linguistic theory they are based upon. In the literature a distinction between *empirical models*, *probabilistic models*, and *language models* is often made, which is rooted in the query-oriented understanding of retrieval tasks but also has historical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

6 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.



http://www.uni-weimar.de/medien/webis/research/projects/retrieval-models

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

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 as embedded in the Euclidean space (see above). Key challenge for 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 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 into 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 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 an effective cross-language retrieval approach, the so-called CL-ESA model. In contrast to most retrieval models, the Suffix Tree Model represents a document d not as a vector of index terms but as a compressed trie containing all suffixes (i.e., suffixes of all lengths) of a text d. As a consequence, the collocation information of d is preserved, which may render the model superior for particular retrieval tasks (Meyer zu Eißen et al, 2005).

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 (Blei et al, 2003). 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. More recently, Le and Mikolov (2014) introduced Paragraph Vector, also known as the Doc2Vec model, which learns continuous distributed vector representations for documents using a neural network classifier.

7 Relevance Computation

Despite the large variety of retrieval models that have been developed so far, computing the relevance ρ of a document for a query usually boils down to a multiplication of two feature vectors: (1) a feature vector $\mathbf{q}$ representing query q, and (2) a feature vector $\mathbf{d}$ representing document d:

$$\rho(q,d) = \mathbf{q}^T \cdot \mathbf{d} = \sum_{i=1}^{|\mathbf{q}|} \mathbf{q}_i \cdot \mathbf{d}_i$$

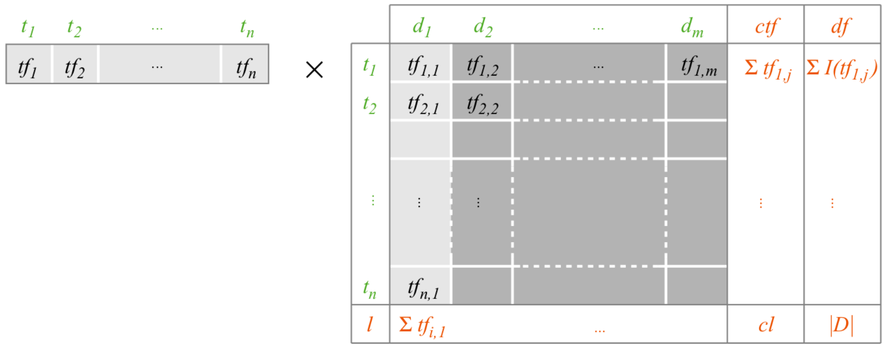

t_1 t_2 ... t_n
tf_1 tf_2 ... tf_n
×
d_1 d_2 ... d_m
ctf
df
t_1 $tf_{1,1}$ $tf_{1,2}$ ... $tf_{1,m}$ $\Sigma\, tf_{1,j}$ $\Sigma\, I(tf_{1,j})$
t_2 $tf_{2,1}$ $tf_{2,2}$
t_n $tf_{n,1}$
l $\Sigma\, tf_{i,1}$ ... cl $|D|$

Fig. 2 Basic model for relevance computation: the term frequency vector of a query (left) is combined with the term frequency vectors of the documents d_j from a document collection D (shaded matrix on the right).

What distinguishes retrieval models from each other is the feature set that they employ for representing queries and documents, as well as the computation rule used to calculate the respective feature weights. In the following, the feature sets and the computation rules for four retrieval models are outlined, starting from the basic tf-Model to the more sophisticated models $tf \cdot idf$, BM25, and ESA.

tf-Model . The tf-Model (*term frequency* model) is a variant of the Vector Space Model that uses the vocabulary of the given document collection D as feature set. The dimension i of a feature vector is associated with a specific term t_i that occurs in D. As feature weight, the frequency tf by which t_i appears in a specific query or document is taken:

$$\mathbf{q}_i^{tf} = tf(t_i, q)$$
$$\mathbf{d}_i^{tf} = tf(t_i, d)$$

Stacking the tf feature vectors of all documents $d \in D$ as columns into a matrix gives the so called *term-document-matrix*, a data structure from which many retrieval models can be derived. The term-document-matrix (along with additional statistics used later on, shown in red) is depicted in Figure 2 as shaded area on the right-hand side. On the left-hand side, a canonical query vector is shown. Multiplying the query vector with the term-document-matrix results in a vector containing the relevance scores for all documents in D. A particular property of the tf-Model is that the relevance computation for a document d is independent of the collection D from which d is taken. However, the model has certain weaknesses that the following models try to alleviate.

$tf \cdot idf$-Model . An extension of the tf-Model is the $tf \cdot idf$-Model , where idf stands for *inverse document frequency*. The linguistic intuition behind this extension is that the occurrence of a rare query term in a document is a better indicator for relevance than the occurrence of a frequent term. Considering the query “math

for computer science" as an example, the occurrence of the term "for" in a document provides, in comparison to the occurrence of "math" or "computer science", only little evidence about the relevance of a document—a fact which is not exploited in the *tf*-Model . A document containing the query term "for" ten times is considered as relevant as a document containing the query term "math" ten times. To address this deficit, the $tf \cdot idf$-Model incorporates the document frequency *df* of a term into the feature weight computation for documents. The document frequency denotes the number of documents in D that contain a term. In Figure 2, *df* is illustrated as an additional column of the term-document-matrix. To formalize the computation of *df* the indicator function $I(\cdot)$ is used, which yields 1 in case $tf(t_i, d_j) > 0$ and 0 otherwise. Since the influence of a term should decrease with its document frequency, an *inverse document frequency* factor is added to the *tf* feature weight computation. Note that several variations for this factor have been proposed. The original formula by Salton et al (1975) reads as follows:

$$\mathbf{d}_i^{tfidf} = tf(t_i, d) \cdot \log \frac{|D|}{df_i}$$

BM25-Model . The BM25-Model is a further advancement of the $tf \cdot idf$-Model . The development of the model was driven by the observation that the $tf \cdot idf$-Model (1) is biased towards long documents, and (2) that it insufficiently favors documents containing all query terms—compared to documents containing only a subset of the query terms. To account for the first observation, the BM25-Model introduces a length normalization factor to the feature weight computation. The length of a document is considered as the sum of its term frequencies. In Figure 2, the document length l is illustrated as an additional row of the term-document-matrix. The sum over all $l \in D$ gives the overall collection length cl. The idea of the length normalization factor is to calculate the average length $\hat{l}$ of the documents, $\hat{l} = cl/|D|$, and to penalize documents that are longer than the average, while rewarding shorter documents. To account for the second observation, a term frequency normalization factor is introduced. Given the above example query "math for computer science", the goal of this factor is to consider a document containing both "math" and "computer science" once as more relevant than a document containing one of the terms twice, even if the terms have equal document frequency. The BM25 approach applies a "logarithmic-shaped" function to the term frequency value in order to limit the contribution of a single term to the overall relevance score. The general form of this function is $\frac{tf}{tf+c}$, where c is a constant. The final BM25 formula for the computation of feature weights, which incorporates both normalization factors into a single expression, is the product of extensive empirical evaluation efforts:

$$\mathbf{d}_i^{bm25} = \frac{tf(t_i, d) \cdot (k_1 + 1)}{tf(t_i, d) + k_1 \cdot (1 - b + b \cdot \frac{l(d)}{\hat{l}})} \cdot \log \frac{|D| - df_i + 0.5}{df_i + 0.5}$$

For the two parameters of the function, values of $k_1 = [1.2, 2.0]$ and $b = 0.75$ are considered standard choices. The two normalization factors (length normalization and term frequency normalization) are balanced by the parameter b. The last factor of the formula is the BM25 variant of the inverse document frequency.

ESA-Model. The ESA-Model represents a class of retrieval models that do not employ terms as features but concepts or topics (hence called "topic models"). Other retrieval models of this kind are LSI and LDA. Topic models aim to further improve the assessment of relevance by taking the semantic relatedness of terms into account. The intuition is that if a document contains terms related to the query terms, like "statistics" or "calculus" which are related to the query term "math", or "programming" and "algorithm" which are related to "computer science", the relevance of this document should be raised. To operationalize this idea, topic models represent queries and documents by a feature vector of topics and provide a means to compute the relevance of a topic for a query or document. In the case of ESA, topics are drawn randomly from the set of Wikipedia articles, and the $tf \cdot idf$-Model is employed to represent each drawn article a as a term based feature vector $\mathbf{a}$. The assumption underlying this approach is that each feature vector $\mathbf{a}$ will contain high $tf \cdot idf$ scores for semantically related terms. To compute the relevance of a topic for a document or query, the cosine similarity between the $tf \cdot idf$ representation of the topic and the $tf \cdot idf$ representation of the query or document is used:

$$\mathbf{q}_i^{esa} = cos(\mathbf{q}^{tfidf},\ \mathbf{a}_i^{tfidf})$$
$$\mathbf{d}_i^{esa} = cos(\mathbf{d}^{tfidf},\ \mathbf{a}_i^{tfidf})$$

8 Key Applications

The key application of retrieval models is to provide keyword-based search capabilities over large collections of natural language text such as digital libraries or the World Wide Web. In many practical settings, the documents of a collection are not completely unstructured but come along with designated meta data such as document titles, abstracts, or markup in the text as in the case of web pages. By taking this additional information into account, e.g. through boosting the relevance score of documents that contain the query terms in the title, the quality of the search can often be improved significantly over the use of standard retrieval models. In the field of Web search, probably the most prominent approach in this respect is the PageRank score (Page et al, 1999), which exploits the hyperlink graph of the Web for the relevance assessment of web pages.

9 Future Directions

Classical retrieval models provide the formal means of satisfying a user's information need (typically a query) against a large document collection such as the Web.

These models can be seen as heuristics that operationalize the probability ranking principle mentioned above. Regarding future directions, a new generation of retrieval models may be capable to support information needs of the following kind: “Given a hypothesis, what is the document that provides the strongest arguments to support or attack the hypothesis?”

Obviously, the implied kind of relevance judgments cannot be made based on the classical retrieval models, as these models do not capture argument structure. In fact, so far the question of how to exploit argument structure for retrieval purposes has hardly been raised, but the research community has picked up this exciting direction (Gurevych et al, 2016).

10 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

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of Machine Learning Research 3:993–1022

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: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, pp 1606–1611

Gurevych I, Hovy E, Slonim N, Stein B (2016) Debating Technologies (Dagstuhl Seminar 15512). Dagstuhl Reports 5(12):18–46, DOI http://dx.doi.org/10.4230/DagRep.5.12.18, URL http://drops.dagstuhl.de/opus/volltexte/2016/5803

Le QV, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning (ICML2014), pp 1188–1196

Meyer zu Eißen S, Stein B, Potthast M (2005) The Suffix Tree Document Model Revisited. In: Tochtermann K, Maurer H (eds) 5th International Conference on Knowledge Management (I-KNOW 05), Know-Center, Graz, Austria, Journal of Universal Computer Science, pp 596–603

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web.

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, DOI http://doi.acm.org/10.1145/290941.291008

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