Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
Set-theoretic	Boolean	Represents documents by a set of index terms. Each term is viewed as Boolean variable, valued as true if the term is present in the document, false otherwise. Terms are not weighted. Queries are specified as arbitrary Boolean expressions, which are formed by linking terms with the logical operators AND, OR, and NOT. Documents are not ranked.
	Fuzzy Set	Based on fuzzy set theory, which accommodates the notion of a *degree of membership* of terms in documents. The degree of membership varies from 0 to 1. The logical operators AND, OR, and NOT are suitably redefined to reflect degree of membership. The fuzzy set model considers the correlations among the terms. However, it does not consider the frequency of occurrence of terms in documents. Not a widely studies and used model.
	Extended Boolean	Extends the Boolean model with functionality for partial matching and term weighting. It combines the characteristics of the vector space model with Boolean query formulation. Uses a generalized scalar product for computing similarity between a document and query. The well-known L_p norm defined for an $n-$dimensional vector is used for the scalar product. The interpretation of a query is altered by the value chosen for p. When $p = 1$, the model behaves like a Vector Space model; when $p =$ inf and the query terms are all equally weighted, behaves like the fuzzy set model; when $p =$ inf and the query terms are not weighted, behaves like the Boolean model.
Algebraic	Vector Space	Both documents and queries are represented as vectors, in an $n-$dimensional space spanned by a set of orthonormal term vectors. The relevance of a document to a query is based on a similarity measure, which is the scalar product of the document and query vectors. It was the most widely used model until end of last century.

Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
	Latent Semantic Indexing	Latent Semantic Analysis (LSA) is a technique in Natural Language Processing (NLP) for analyzing *distributional semantics* – relationships between a set of documents and the terms contained in the documents. LSA is based on the *distributional hypothesis* – words that are close in meaning will appear in similar documents. LSA aims to estimate the pattern of word usage across documents. An enhanced term-document incidence matrix with term frequency counts is constructed, and the number of rows is reduced using Singular Value Decomposition (SVD). In the resulting low-dimensional space, each row is a term vector and the similarity between the terms is computed as the cosine of the angle between their term vectors. An IR model that uses this latent semantic structure is called latent semantic indexing (LSI). LSA assumes that terms and documents form a joint Gaussian model, and this does not match the observed data. An alternative, Probabilistic Latent Semantic Analysis (PLSA) and its application to IR called the Probabilistic Latent Semantic Indexing (PLSI), has shown to perform better than LSA/LSI. An advantage of this technique is that queries can retrieve documents even when the query and documents have no common words.

Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
	Generalized Vector Space	The Generalized Vector Space Model (GVSM) aims to overcome the pairwise orthogonality assumption of the Vector Space Model (VSM). GVSM overcomes the pairwise orthogonality assumption by introducing term-to-term correlations. Term vectors are represented in terms of smaller components called *minterms*, which are binary indicators of all patterns of occurrence of terms in documents. Minterms are represented as vector, and a minterm vector represents one type of co-occurrence of terms in a document. For example, a minterm vector corresponding to terms t_1 and t_2 points to only documents in which t_1 and t_2 co-occur. For n terms, there will be 2^n minterm vectors. Pairwise orthogonal vectors associated with the minterms comprise the *basis vectors* for GVSM. The similarity between the query and document is calculated in the space of minterm vectors. The advantage of GVSM is that it considers correlations among terms. The computational cost is high for large document collections.
	Neural Network	Neural network models for IR use Machine Learning (ML) based approaches. They are an evolution of traditional *learning to rank* models. While the learning to rank models employ ML algorithm using hand-crafted IR features, neural models learn representations of language directly from raw text. However, neural models require large-scale training data for their development. Neural approaches can be *shallow* or *deep* depending on the number of layers in the neural architecture. Some neural IR models base relevance on lexical matching only, while others are able to extract relevance from related terms as well using semantic matching. Current models seem to not perform well for queries that involve *rare* terms. Deep neural network models for IR is an active area of current IR research.

Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
Probabilistic	Classical	Probabilistic models are a family of models based on the *Probability Ranking Principle* (PRP). These models consider documents and queries as observations from random variables. They rank documents by estimating the probability of relevance for document-query pairs. The models differ based on the assumptions they make. The classical model makes *term independence assumptions* – terms occur in documents independent of each other. Documents are ranked based on estimated probability of relevance of documents to query. Binary Independence Model (BIM) is a classical model which represents both documents and queries as binary vectors. Okapi BM25 models extends BIM by incorporating term frequency and document length normalization. BM25 has been the popular and widely used IR model.
	Statistical Language Model	A statistical language model is a probability distribution over all word sequences in the language. The language model estimates the relive likelihood of all word sequences. In the language model approach to IR, there is a language model associated with each document. The relevance of a document to a query is the probability that the query is most likely has been generated by the language model of the document. Use this probability to rank documents. Language model based approaches to IR are popular and widely used.

Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
	Divergence from Randomness	It is a generalization of Harter's 2-Poisson indexing-model. The 2-Poisson model is based on the hypothesis that informative terms occur more frequently in a set of documents (referred to as *elite* set) than in the rest of the documents. For the terms which do not possess elite documents, their term frequencies follow a random distribution. The model is based on a simple idea: the more the divergence of a frequency of a term t in document d from its collection frequency, the more the information carried by the term t in document d. A Divergence from Randomness (DFR) model is created by instantiating the three components of the framework: selecting a basic randomness model, applying the first normalization, and normalizing the term frequencies. Combination of these choices lead to different DFR models. Poisson model with Laplace after-effect and normalization 2 (aka PL2 model) is a well-known DFR model.
	Probabilistic Inference	probabilistic inference models are based on epistemological view of probability. Under this interpretation, probability is viewed as a degree of belief an individual places on the uncertainty of a particular situation. The model assumes that there exists an ideal *concept space* (aka domain of reference), and the elements of this space are *elementary concepts*. A proposition is a subset of the concept space. Documents, terms contained in the documents, and user queries are all represented as propositions in the concept space. A probability function p is defined on the concept space. For a document d, $p(d)$ is interpreted as the degree to which the concept space is covered by the knowledge contained in d. Likewise, for a given query q and document d, $p(q \cap d)$ quantifies the degree to which the knowledge common to both q and d is covered by the concept space. These probabilities are used to rank documents with respect to a query.

Table 1: A categorization of IR models

IR Model Class	Subclasses	Distinguishing Characteristics
Axiomatic		Axiomatic approach provides a framework for developing new retrieval models. The approach is based on formally defined retrieval constraints at the term level. A retrieval model is found by searching in a space of candidate retrieval functions for one that satisfies a chosen set of retrieval constraints. Several new retrieval functions have been derived using the axiomatic approach. Experimental results show that the derived retrieval functions are more robust, and less sensitive to parameter settings than the existing retrieval functions with comparable optimal performance.