



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

A Study and Analysis of Information Retrieval Models

S.Ruban¹, Dr.S.Behin Sam², Lenita Veleza Serrao³, Harshitha³

Assistant Professor, Dept. of MCA, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India¹

Assistant Professor, Dept. of Computer Science, Rajeswari Vedachalam College, Tamil Nadu, India²

Student, Dept. of MCA, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India³

ABSTRACT: Information Retrieval deals with the representation, storage, organization of and access to Information items. The evaluation of an information retrieval system is the process of assessing how well a system meets the information needs of its users. Information retrieval (IR) systems use a simpler data model than database systems. Information organized as a collection of documents. The effective retrieval of relevant information is directly affected both by the user task and by the logical view of the documents adopted by the retrieval system. Information Retrieval (IR) is the method that contract with retrieval of unstructured data mainly textual documents, in response to a query or topic statement, which may itself be unstructured. Any Information Retrieval system is based on Information Retrieval models. In this paper we will be discussing about the different Information Retrieval models that has been scheduled and will estimate some models based on certain evaluation measures.

KEYWORDS: IR Process, Query, Information Retrieval, Information Retrieval Model, Ranking, Indexing, Retrieval Process.

I. INTRODUCTION

Information Retrieval deals with the representation, storage, organization of and access to Information items. Information retrieval has more than one definition. Information Retrieval can also said to be tracing and recovery of specific information from stored data. It also said to be that it is the activity of obtaining information resources relevant to an information need from a collection of information resources, searches can be based on metadata or on full-text(or other content based) indexing.

An information retrieval begins when user enters a query into the system. Query does not uniquely identify a single object in the collection but are instead represented in the system by document surrogates or metadata. The systematic storage and recovery of data as from a file card catalogue or the memory bank of a computer. Information retrieval consists of 4 main stages identifying the exact subject of the search, locating this subject in a guide which refers the searcher to one or more documents, locating the documents, locating the required information in the documents. Entity is identified as object which is represented by information in a database. Queries which are sent by user are matched against database information. The evaluation of an information retrieval system is the process of assessing how well a system meets the information needs of its users. Information retrieval (IR) systems use a simpler data model than database systems. Information organized as a collection of documents

Information retrieval locates relevant documents, on the basis of user input such as keywords or example documents, e.g., find documents containing the words “database systems”. It can be used even on textual descriptions provided with no textual data such as images. Web search engines are the most familiar example of IR systems. We can give some differences between database systems and Information Retrieval:

- IR systems don't deal with transactional updates (including concurrency control and recovery).
- Database systems deal with structured data, with schemas that define the data organization.
- IR systems deal with some querying issues not generally addressed by systems.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

II. INFORMATION RETRIEVAL PROCESS

The retrieval process is interpreted in terms of component sub processes which interact with one another. To describe the retrieval process, we use simple and generic software architecture as shown in Figure 1. First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following:

- (a) The documents to be used,
- (b) The operations to be performed on the text, and
- (c) The text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them. Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted file as indicated in Figure . The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times.

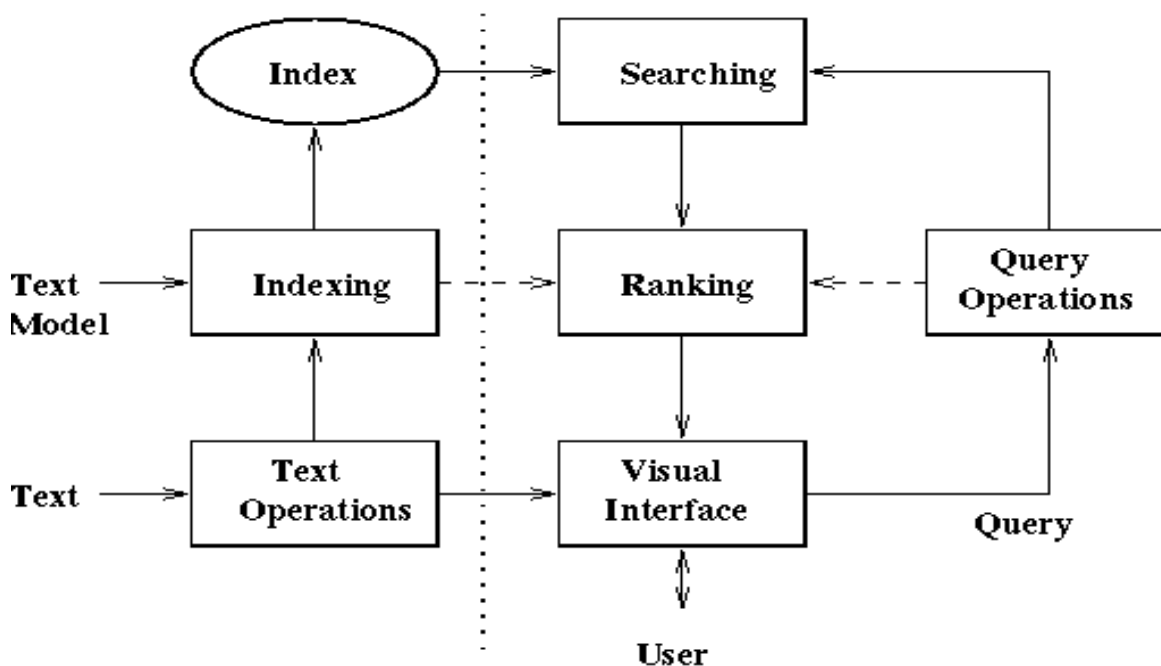


Figure 1: Informational retrieval process

Given that the document database is indexed, the retrieval process can be initiated. The user first specifies a *user need* which is then parsed and transformed by the same text operations applied to the text. Then, *query operations* might be applied before the actual *query*, which provides a system representation for the user need, is generated. The query is then processed to obtain the *retrieved documents*. Fast query processing is made possible by the index structure previously built.


Before been sent to the user, the retrieved documents are ranked according to a *likelihood* of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a *user feedback* cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation of the real user need.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

The small numbers outside the lower right corner of various boxes in Figure  indicate the chapters in this book which discuss the respective sub processes in detail. A brief introduction to each of these chapters can be found in section Consider now the user interfaces available with current information retrieval systems (including Web search engines and Web browsers). We first notice that the user almost never declares his information need. Instead, he is required to provide a direct representation for the query that the system will execute. Since most users have no knowledge of text and query operations, the query they provide is frequently inadequate. Therefore, it is not surprising to observe that poorly formulated queries lead to poor retrieval (as happens so often on the Web).

III. INFORMATION RETRIEVAL MODELS

Any Information Retrieval system is based on Information Retrieval models. An Information Retrieval Model is a quadruple $\{D, Q, F, R(q_i, d_j)\}$ where

- i) D is a set made of logical aspect (or representations) for the documents in the group of collection.
- ii) Q is a set made of logical design (or representations) for the end user information needs.
- iii) F is a Figure for classical document representations, queries, and their relationships.
- iv) $R(q_i, d_j)$ is a ranking function which affiliate a real number with a query $q_i \in Q$ and a document representation $d_j \in D$.

Such ranking describes an ordering among the documents with concern to the query q_i . A model of information retrieval deal as a blueprint which is used to appliance an actual information retrieval system.

Forms of Information Retrieval Models:

1. Exact Match Models :
 - Boolean model.
 - Region model.
2. Vector space model
3. Probabilistic Approaches
 - Probabilistic indexing model.
 - Probabilistic retrieval model.
 - 2-poisson model
 - Bayesian network model.
 - Language model
4. Set Theoretic models
 - Fuzzy set model.
 - Extended Boolean model.
5. Algebraic models
 - Generalized vector space model.
 - Latent Semantic indexing model.
 - Neural Network model

IV. THE BOOLEAN INFORMATION RETRIEVAL MODEL

Based on set theory and Boolean algebra the Boolean Information Retrieval model is a simple model. This model considers that index terms are present or absent in a document. Hence the index term weights are assumed to be all in binary i.e. $\{0,1\}$. A query q is composed of index terms linked by three connectives: NOT, AND, OR. For instance, the query STUDIES and SPORTS will produce the set of documents that are indexed both with the term STUDIES and the term SPORTS which will be the intersection of both sets. The OR operator that combines the terms will define a document set which is greater than or equal to the document sets of the single terms. Hence the query STUDIES OR SPORTS will produce the set of documents that are indexed either with the term STUDIES or with the term SPORTS or both which will be union of both sets. In Boolean model there is an advantage which gives users a sense of control over

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

the system. Hence a document will be retrieved immediately whenever a query is given. Hence it is clear which operators will produce a bigger set or smaller set depending on the resulting document.

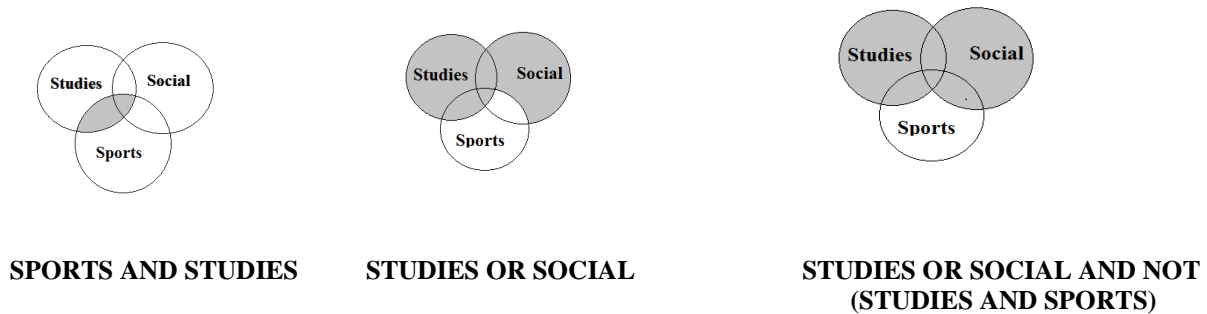


Figure 2: Boolean combinations of sets visualised as Venn diagrams

V. THE VECTOR INFORMATION RETRIEVAL MODEL

Since Boolean matching and binary weights is too limiting, the vector model is proposed in which partial matching is possible. Based on Luhn's similarity criterion a model was suggested by Gerald Salton and his colleagues that has a stronger theoretical motivation. According to them the index representations and the query are vectors embedded in a high dimensional Euclidean space where in each term are assigned a separate dimension. Generally the similarity measure is the cosine of the angle that separates the two vectors d and q . The cosine of an angle is 0 if the vectors are orthogonal in the multidimensional space and 1 if the angle is 0 degrees. The cosine formula is given by:

$$\text{score}(\vec{d}, \vec{q}) = \frac{\sum_{k=1}^m d_k \cdot q_k}{\sqrt{\sum_{k=1}^m (d_k)^2} \cdot \sqrt{\sum_{k=1}^m (q_k)^2}}$$

Figure 3 visualizes an example document vector and an example query vector in the space that is spanned by the three terms STUDIES, SPORTS and SOCIAL.

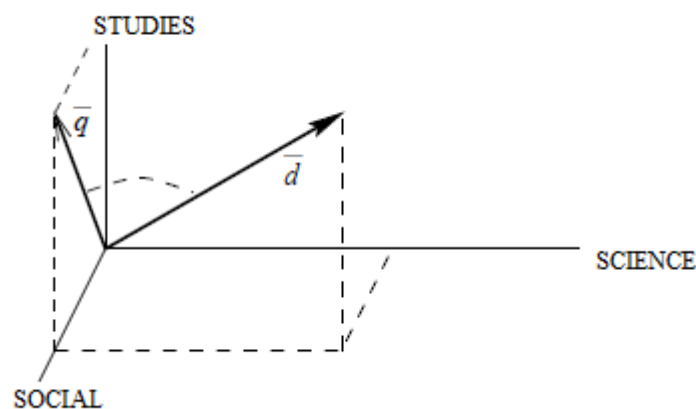


Figure 3: A query and document representation in the vector space model

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Measuring the cosine of the angle between vectors is equivalent with normalizing the vectors to unit length and taking the vector inner product. The formula then becomes:

$$\text{score}(\vec{d}, \vec{q}) = \sum_{k=1}^m n(d_k) \cdot n(q_k) \quad \text{where } n(v_k) = \frac{v_k}{\sqrt{\sum_{k=1}^m (v_k)^2}}$$

VI. THE PROBABILISTIC INDEXING MODEL

Bill Maron and Larry Kuhns introduced the probabilistic indexing model. They did not target automatic indexing by information retrieval systems. They proposed a human indexer since manual indexing was still guiding the field, which runs through the various index items T that possibly apply to a document D and assigns a probability P(T/D) to a term given a document instead of making a yes/no decision for each term. Hence each document ends up with a set of possible index terms weighted by P(T/D), where P(T/D) is the probability that if a user needs information of the kind contained in document D, he/she will formulate a query by using T.

Using Bayes' rule, i.e.,

$$P(D/T) = \frac{P(T/D)P(D)}{P(T)}$$

Later they considered to rank the documents by P(D/T), i.e. the probability that the document D is formulated a query by using the term T. P(T) in the denominator of the right-hand side is constant for any given query term T, and consequently documents might be ranked by P(T/D)/P(D) which is a quantity proportional to the value of P(D/T).

In the formula, P(D) is the a-priori probability of relevance of document D.

VII. THE PROBABILISTIC RETRIEVAL MODEL

Stephen Robertson and Karen Sparck-Jones introduced the probabilistic retrieval model based on reasoning. They proposed to rank documents by P(R/D), i.e. the probability of relevance R given the document's content description D. Here in this model D is a vector of binary components each component representing a term whereas in the probabilistic indexing model section D was the "relevant document".

The probability P(R/D) in the probabilistic retrieval model has to be interpreted as follows: there can be several, say 10, documents that are represented by the same D. If 9 of them are relevant, then P(R/D) = 0:9.

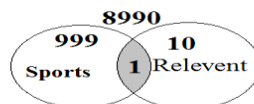


Figure: 3 Venn Diagram with query term sports

To make this work in practice, we use Bayes' rule on the probability odds P(R/D)/P(R/D), where R denotes irrelevance. The odds allow us to ignore P(D) in the computation while still providing a ranking by the probability of relevance. Additionally, we assume independence between terms given relevance.

$$\frac{P(R|D)}{P(\bar{R}|D)} = \frac{P(D|R)P(R)}{P(D|\bar{R})P(\bar{R})} = \frac{\prod_k P(D_k|R)P(R)}{\prod_k P(D_k|\bar{R})P(\bar{R})}$$

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Here, D_k denotes the k th component (term) in the document vector. A more convenient implementation of probabilistic retrieval uses the following three order preserving transformations. First, the documents are ranked by sums of logarithmic odds, instead of the odds themselves. Second, the a priori odds of relevance $P(R)/P(\bar{R})$ is ignored. Third, we subtract $\sum_k \log(P(D_k=0|R)/P(D_k=1|R))$, i.e., the score of the empty document, from all document scores. This way, the sum over all terms, which might be millions of terms, only includes non-zero values for terms that are present in the document.

$$\text{matching-score}(D) = \sum_{k \in \text{matching terms}} \log \frac{P(D_k=1|R) P(D_k=0|\bar{R})}{P(D_k=1|\bar{R}) P(D_k=0|R)}$$

VII. FUZZY SET MODEL

Query term is matched to a document which is approximate or vague. Using a Fuzzy framework this vagueness can be modelled as follows. Every query term defines a fuzzy set and each document has a degree of membership in this set. Based on fuzzy theory this interpretation provides the foundation for many IR models. Here we discuss the model proposed by Ogawa, Morita and Kobayashi. Fuzzy set theory deals with the representation of classes whose boundaries are not well defined. The main idea here is to introduce the notion of a degree of membership associated with the elements of the class which varies from 0 to 1 and allows modelling the notion of marginal membership. Thus, membership is now a gradual notion, contrary to the crispy notion enforced by classic Boolean logic. A fuzzy subset A of a universe of discourse U is characterized by a membership function

$$\mu_A : U \rightarrow [0, 1]$$

This function associates with each element u of U a number $\mu_A(u)$ in the interval $[0, 1]$. The three most commonly used operations on fuzzy sets are:

- The complement of a fuzzy set.
- The union of two or more fuzzy sets.
- The intersection of two or more fuzzy sets.

Fuzzy sets are modelled based on a thesaurus, which defines term relationships. A Thesaurus can be constructed by defining a term-term correlation matrix C . Each element of C defines a normalized correlation factor $c_{i,l}$ between two terms k_i and k_l .

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

Where n_i : number of docs which contain k_i , n_l : number of docs which contain k_l and $n_{i,l}$: number of docs which contain both k_i and k_l .

We can use the term correlation matrix C to associate a fuzzy set with each index term k_i . In this fuzzy set, a document d_j has a degree of membership $\mu_{i,j}$ given by

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$$

The above expression computes an algebraic sum over all terms in d_j . A document d_j belongs to the fuzzy set associated with k_i , if its own terms are associated with k_i . If d_j contains a term k_l which is closely related to k_i , (i.e., $c_{i,l} \sim 1$) then $(\mu_{i,j} \sim 1)$ and k_i is a good fuzzy index for d_j .

VIII. EVALUATION FEATURES

Each of the Information Retrieval models can be characterized by the features they implement as well as the performance they have in different scenarios. We define some common features that can be used to describe each information retrieval model based on the functionalities and characteristics they possess.

Exact Matching : Indicates the way whether the query exactly matches the document.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Partial Matching : Indicates the way whether the keyword in the query partially matches the document.

Relevance : It denotes how well a retrieved document or set of documents meets the information need of the user.

Indexing : In order to be able of making efficient searches over the document collection. It is necessary to have the data stored in specially designed data structures.

Ranking : It is in charge of sorting the results, based on heuristics that try to determine which results satisfy better the user need.

Smoothing: It allows fine tuning the ranking to improve the results.

Probability : It is a measure of the expectation that an event will occur or a statement is true. Probabilities are given a value between 0 (will not occur) and 1 (will occur). The higher the probability of an event, the more certain we are that the event will occur.

Similarity : It is a measure to evaluate how a query matches a document.

Thesaurus : It is a reference work that lists words grouped together according to similarity of meaning.

IX. EVALUATION

Retrieval Model	Exact	Partial	Relevance	Indexes	Rankings	Smoothing	Probability	Similarity	Thesaurus
Boolean Model	Yes	No	No	Yes	No	No	No	yes	No
Vector Space Model	Yes	Yes	Yes	No	No	No	No	Yes	No
Probabilistic Indexing Model	Yes	Yes	Yes	Yes (Manual)	No	No	Yes	No	No
Probabilistic Retrieval Model	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Fuzzy Set Model	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes

Figure :4 Characterization of Retrieval Models

X. CONCLUSION

There is no such dominating model or theory of information retrieval in the case of information retrieval system. We have seen some of the information retrieval models here. Some of the models may work for some of the applications and other models may work for other applications. This chapter tells us about the consequences assumptions. User can read and once he or she is aware of the consequences assumptions, it is possible for them to pick a information retrieval model that is adequate in new stages. It is possible to bring the improvements in information retrieval models by announcing newer models.

REFERENCES

- [1] Ricardo Baeza-Yates, Ribeiro-Neto, Modern Information Retrieval, ACM press.
- [2] Rajendra Akerkar, Pawan Lingras, Building an Intelligent Web- Theory and Practice, Jones and Bartlett Publishers.
- [3] Goker A, Davies.J , Information Retrieval: Searching in the 21st Century, John Wiley and sons, Nov 2009.
- [4] Salton G.M.McGill, Introduction to Modern Information Retrieval, McGraw Hill.
- [5] Turtle H, W. croft Evaluation of an inference network-based retrieval model, ACM Transactions on Information Systems.
- [6] Sparck Jones,K.S. Walker, S.Robertson. A probabilistic model of information retrieval: Development and Comparative experiments(part 1 and 2). Information Processing and Management. 36(6), 779-840.
- [7] Y.Ogawa, T. Morita, K.Kobayashi, A fuzzy document retrieval System using the key connection matrix and a learning method, fuzzy sets and systems 39:163-179.
- [8] G.A Miller, 1990, Special Issue, Wordnet: An on-line lexical database, International journal of Lexicography, 3(4).
- [9] Ellen M. Voorhees, 1994, Query Expansion using Lexical-semantic relations, In proceedings of the 17th ACM-SIGIR Conference, pages 61-69.
- [10] Salton G & Buckley C (1990). Improving Retrieval performance by relevance feedback, Journal of the American society for Information Science.