"Trends and issues in Modern Information Retrieval"

Arman Rasool Faridi[1] Aasim Zafar[2]

[1]Department of Computer Science, Aligarh Muslim University, Aligarh, arman.faridi@gmail.com

[2]Department of Computer Science, Aligarh Muslim University, Aligarh, aasimzafar@gmail.com

Abstract – This paper attempts to present an overview of the modern information retrieval models. Firstly, the classical IR models have been discussed and then their improved versions commonly called as alternative or modern information retrieval models have been briefly presented. In the end the main issues and challenges in modern information retrieval techniques have been summarized.

Keywords: Information Retrieval, Boolean Model, Vector Space Model, Probabilistic Model, Issues in Information Retrieval

----------♦----------

1. INTRODUCTION

Information retrieval is about retrieving information relevant to the user on the basis of a query. It is now considered as an important branch of computer science with a goal to provide effective methods for satisfying information needs. It is distinguished from database retrieval as the representation of information is usually more loosely structured compared to the rigid table-based organization of a database, and that the information need is often not completely specified. However, the distinction between these two fields is becoming more obscure as more structure is used in documents.

The scope of information retrieval (IR) is as broad as information itself. Early IR research (during the mid-1950s to the late 1970s), were mainly focused on text-based information retrieval. In modern times, however, the scope of retrieval algorithms has been widened to retrieve the information available in video, digital photos, scanned and on-line handwriting, genetic data, music, audio clips, and hypertext formats. Further, cross-lingual IR techniques are available to handle hundreds of different human languages.

There is a general formulation of information retrieval that all of these applications share. A retrieval algorithm is given a *query* generated by a user that represents their information need. In the case of text, this query consists of a series of words, along with possibly a set of relations between them. It is assumed that the information to be found resides in a *collection* which consists of a set of *documents*. Here the term document is very general and refers to a basic unit of information that could be a Web page, image, audio clip, and so on. Given the *query*, the retrieval algorithm then scores the *documents* in the *collection*, ranking them according to some measure of how well the query terms and relations are matched by information in the document. For text, the relations most often used between terms are co-occurrence or proximity constraints. Traditional relevance also relies on the frequency with which terms occur in a document, and how unusual the terms are in the collection.

2. MODERN INFORMATION RETRIEVAL TECHNIQUES

Historically, the term Information Retrieval was coined by Calvin Mooers in 1950. Since then several approaches of information retrieval have been discussed. Some of these are termed as Classical IR models. Over the period of time several alternative models based on these classical models have been proposed to overcome the weaknesses, widen the scope and improve the performances of existing IR systems. In the following section, a brief overview of these IR models is presented.

2.1. Boolean Model

In the Boolean retrieval model of information retrieval, each document is associated with a set of keywords or index terms. The user can pose any query which is in the form of

a boolean expression of terms, that is, in which terms are combined with the operators AND, OR, and NOT. This model views each document as just a set of words. The main advantage of this model is that it is easy to implement and also computationally efficient (Baeza-Yates and B. Ribeiro-Neto, 1999).

A strict boolean expression over terms with an unordered results set is too limited for many of the information needs of the users. To cater the information needs of such users, extended Boolean retrieval models were implemented by incorporating additional operators such as term proximity operators. A proximity operator is a way of specifying that two terms in a query must occur close to each other in a document, where closeness may be measured by limiting the allowed number of intervening words or by reference to a structural unit such as a sentence or paragraph.

2.1.1 Extended Boolean Model

In the extended Boolean model, a document has a weight associated with each index term. This document weight is a measure of the degree to which the document is characterized by that term. Without loss of generality, it is assumed that document weights for all index terms lie in the range [0, 1]. This is less restrictive than in the standard Boolean model, which limits the values to the extremes of the range, namely 0 and 1.

To retrieve documents relevant to a given query, it is required to calculate the query-document similarity for documents in the collection. The query-document similarity is an attempt to predict the relevance of a document to the query. Three very commonly used models under this category used for calculating similarity are MMM, Paice and P-norm.

2.1.2 Fuzzy Set Model

In Boolean model, the retrieval strategy is based on binary criteria. So, partial matches are not retrieved. Only those documents that exactly match the query are retrieved. Hence, to effectively retrieve from a large set of documents, users must have a good domain knowledge to form good queries. Further, it is not appropriate for handling uncertainty, imprecision, vagueness and inconsistency which are common in information retrieval systems (IRS).

Information retrieval involves two finite crisp sets, a set of recognized index terms, $X = \{x_1, x_2, \ldots\ldots x_n\}$ and a set of relevant documents, $Y = \{y_1, y_2, \ldots, y_n\}$. In fuzzy information retrieval, the relevance of index terms to individual documents is expressed by a fuzzy relation, $R = X \times Y \dashrightarrow [0,1]$, such that membership value $R(x_i, y_j)$ specifies for each $x_i \in X$ and $y_j \in Y$ the grade of relevance of index term x_i to document y_j.

The fuzzy set model has the following advantages over the classical methods.

- The use of fuzzy set theory makes the fuzzy relevance relations and fuzzy thesauri more expressive than their crisp counterparts, resulting in a more realistic construction.
- The retrieved documents are distinguished by their relevance in fuzzy set model, thereby providing more relevant documents to the user.
- This model provides the user with more flexibility while expressing the query.

2.1.3 Set Based

The set-based model functions for computing the similarity between a document and a query. It considers the term set frequency in the document and its scarcity in the document collection. In this approach set theory is used for calculating term weights and for ranking documents. In one of the implementation, the set-based model uses a term weighting scheme based on association rules theory [Agrawal, Imielinski, & Swami, 1993].

2.2. Vector Model

The main problem with Boolean model is its inability to fetch partial matches and the absence of any scoring procedure to rank the retrieved documents. This problem was addressed in the vector based model of Information retrieval.

The Vector Space Model (VSM) has been adopted in information retrieval as a means of coping with inexact representation of documents and queries, and the resulting difficulties in determining the relevance of a document relative to a given query. VSM relies on the assumption that the meaning of a document can be derived from the document's constituent terms. A vector is used to represent each document in collection. Each component of vector reflects a term associated with the document. The value assigned to that component reflects the importance of the term in representing semantics of the document. Each document is represented by a weight vector $\vec{d}_j = (w_{1j}, w_{2j}, \ldots, w_{tj})^T$ where w_{zj} is the weight or importance of the term z in representation of the document $\vec{d}_j$, t is the size of the indexing term set. A collection of d documents is then represented by a term-document matrix with t rows

and *d* columns. Query vector representation is given as, $\vec{q}_i = (q_{1i}, q_{2i}, ..., q_{ti})^T$ where q_{zi} is the weight of term *z* in representation of the query $\vec{q}_i$. A variety of models is available in the literature for weighting the document and query vector elements. The cosine similarity between a document and a query is measured (Salton and McGill, 1983). Since the individual terms and keywords are not adequate discriminations of the semantic content of documents and queries, performance of the VSM suffers from two classical problems of synonymy and polysemy [Berry, Drmac & Jessup, 1991] [Kontosthathis & Pottenger, 2006]. The prevalence of synonymy tends to decrease the recall performance of retrieval systems. Polysemy is one factor for poor precision performance.

2.2.1 Generalized Vector Space Model

Generalized Vector Space Model (GVSM) attempts to extend the standard Vector Space Model (VSM) by embedding additional types of information, besides terms, in the representation of documents.

In VSM, the term-document matrix **A** is assumed to be the term occurrence frequency matrix obtained from automated indexing. However, it ignores term correlations. Use of a co-occurrence matrix can be justified only if the documents and term vectors are assumed to be orthogonal. GVSM proposed by Wong represent term vectors explicitly in terms of chosen set of orthonormal basis vectors [Wong, Ziarko & Wong, 1985]. GVSM modifies VSM by introducing some adhoc schemes for including the important effect of term correlation. The correlation matrix provides a model of the relationships that obtain among the corpus indexing terms. The correlation between any two index terms depends on the number of documents in which two terms appear together. For a term-document matrix **A** of dimension *tXd*, GVSM calculates the term correlation matrix **R** of dimension *tXt* by multiplying **A** with its transpose $\mathbf{A}^T$ matrix. Then GVSM calculates similarity between a query vector and document collection as the dot product between query vector, correlation matrix and term-document matrix.

The major strength of the GVSM derives from the fact that it is theoretically sound and elegant. Furthermore, experimental evaluation of the model on several test collections indicates that the performance is better than that of the VSM.

2.2.2 Latent Semantic Indexing Model

Vector Space Model (VSM) and Generalized Vector Space Model (GVSM) represent documents and queries as vectors in a multidimensional space. This high dimensional data puts great demands on computing resources. In order to overcome these problems, Latent Semantic Indexing (LSI) has been proposed which projects the documents into a lower dimensional space. It takes a set of objects that exist in a high-dimensional space and represents them in a low dimensional space, often in a two-dimensional or three-dimensional space for the purpose of visualization.

LSI assumes that there is some underlying or latent structure in word usage that is partially obscured by variability in word choice. A truncated singular value decomposition (SVD) is used to estimate the structure in word usage across documents. Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD. Performance data shows that these statistically derived vectors are more robust indicators of meaning than individual terms. Latent Semantic Indexing (LSI) attempts to improve GVSM model of term correlations by means of dimensionality reduction. It is stated in IR literature that LSI model is 30% more effective than classical VSM models (Hua, 2006).

2.2.3 Neural Network Model

In neural network models, information is represented as a network of weighted, interconnected nodes. In contrast to traditional information processing methods, neural network models are "self-processing" in that no external program operates on the network: the network literally processes itself, with "intelligent behavior" emerging from the local interactions that occur concurrently between the numerous network components (Reggia & Sutton, 1988).

According to Doszkocs, Riggia and Lin (1990), neural network models in general are fundamentally different from traditional information processing models in at least two ways.

- First they are self-processing. Traditional information processing models typically make use of a passive data structure, which is always manipulated by an active external process/procedure. In contrast, the nodes and links in a neural network are active processing agents. There is typically no external active agent that operates on them. "Intelligent behavior" is a global property of neural network models.

- Second, neural network models exhibit global system behaviors derived from concurrent local interactions on their numerous components. The external process that manipulated the underlying data structures in traditional IR models typically has global access to the entire network/rule set,

and processing is strongly and explicitly sequentialized [Doszkocs, Riggia & Lin, 1990].

2.3. Probabilistic Model

Another classic retrieval method is the probabilistic retrieval, where the probability that a specific document will be judged relevant to a specific query, is based on the assumption that the terms are distributed differently in relevant and non-relevant documents (Belkin and Croft, 1992). The probability formula is usually derived from Bayes' theorem. Probabilistic models offer a principled way of managing the uncertainty that naturally appears in many elements within the area of IR.

2.3.1 Bayesian Networks

Nowadays, the dominant approach for managing probability within the field of Artificial Intelligence is based on the use of Bayesian networks (Jensen, 1996; Pearl, 1988), and these have also been used within IR as extensions of classical probabilistic models.

The model decomposes into two parts: a document collection network and a query network. The document collection network is large, but can be precomputed: it maps from documents to terms to concepts. The concepts are a thesaurus-based expansion of the terms appearing in the document. The query network is relatively small but a new network needs to be built each time a query comes in, and then attached to the document network. The query network maps from query terms, to query subexpressions (built using probabilistic or ``noisy'' versions of AND and OR operators), to the user's information need. The result is a flexible probabilistic network which can generalize various simpler Boolean and probabilistic models.

2.3.2 Inference Network Model

Inference network model attempts to understand the content of documents and queries [Croft, 1987] to infer probable relationships between documents and queries. This model is a variant of probabilistic model to represent documents and information needs. Retrieval is viewed as an evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the probability that a given document matches a query. This model generalizes several current retrieval models and provides a framework within which disparate information retrieval research results can be integrated.

2.3.3 Belief Network Model

The belief network model for information retrieval is derived from probabilistic considerations over a clearly defined sample space. This model is founded on a clearly defined sample space which makes it intuitive. Further it is derived from probabilistic considerations over this sample space which simplifies its understanding. It can also be viewed as an alternative to the inference network model proposed in (Turtle & Croft, 1991).

3. ISSUES AND CHALLENGES IN MIR

The key issues with IR models are selection of search vocabulary, search strategy formulations and information overload. The main issues identified with Boolean IR model may be summarized as:

- Partial matches are not retrieved. Only those documents that exactly match the query are retrieved. Hence, effective retrieval requires the users to have a good domain knowledge to form good queries for a large set of documents (Belkin and Croft, 1992).
- Retrieved documents are not ranked.
- For a given a large set of documents the Boolean model either retrieves too many documents or very few documents.
- The model does not use term weights. If a term occurs only once in a document or several times in a document, it is treated in the same way (Gao et al., 2004).

To resolve the limitation arising due to strict boolean expression over terms, extended Boolean retrieval model was proposed which utilizes additional operators like term proximity operators. This result shows some improvements over the Boolean model.

To overcome the issue of exact match as in Boolean Model, Fuzzy Set Model were proposed which considers the correlations among index terms and degree of relevance between queries and docs can be achieved. But it does not take into account the frequency of a term in a document or a query.

This problem of partial fetch matches and the absence of procedure for ranking the retrieved documents were addressed in the vector based model of Information retrieval but Index terms are considered to be mutually independent. This model does not capture the semantics of the query or the document. Also, the computation cost is fairly high with large collections of documents.

This problem of high computing resources requirement is significantly reduced by LSI by dimensionality reduction.

Although LSI shows improvement over previous models still the following problems were observed:

- Difficulty in finding out similarities between terms.
- A compound term is treated as two terms.
- Ambiguous terms create noise
- Time complexity for SVD (Singular Value Decomposition) in dynamic collections is evident.

Compared to Boolean and Vector Based IR models, probabilistic models offer a principled way of managing the uncertainty that naturally appears in many elements within the area of IR. But it incorporates too many assumptions which results in various side effects which makes it difficult to implement in large scale real life practical situations. The field of IR is still challenging to information scientists and attracts many researchers.

4. CONCLUSION

The classical IR models like Boolean and Vector space models make implicit assumptions about relevance and text representation, which affect the design and effectiveness of ranking algorithms. The retrieval models are validated empirically, rather than theoretically (Croft, Metzler, Strohman 2010). Over the years, alternative modeling paradigms for each type of classic model have been proposed with their own set of advantaged and disadvantages. Based on set theoretic models, alternatives like fuzzy and extended Boolean models are distinguished. In the same way for algebraic models alternatives like generalized vector, latent semantic indexing and the neural network models are distinguished. Regarding alternative probabilistic models alternatives like inference network and the belief network models are distinguished. The human activities like information retrieval are hard to formalize and hence though every alternative model has shown some improvement over other still it is associated with its own set of problems.

REFERENCES

Agrawal R., Imielinski T. and Swami A. (1993). “Mining association rules between sets of items in large databases”. In Proceedings of the ACM SIGMOD International Conference Management of Data, pages pp. 207-216, Washington, D.C.

Baeza-Yates R. and Ribeiro-Neto B. (1999). “Modern Information Retrieval”. Addison Wesley.

Belkin N.J. and Croft. W. B. (1992). “Information filtering and information retrieval: two sides of the same coin?”. Communications of the ACM, 35(12): pp. 29–38.

Berry M. W., Drmac Z. and Jessup E. R., “Matrices, Vector Spaces, and Information Retrieval”. SIAM Review, 41, 335–362, 1991.

Croft B., Metzler D. and Strohman T. “*Search Engines: Information Retrieval in Practice”.* Pearson”, USA, 2010.

Croft W. B. “Approaches to intelligent information retrieval”. Information Processing and Management, 23(4):249-254, 1987.

Doszkocs, T. E., Reggia, J. and Lin, X. "Connectionist models and information retrieval". In Annual Review of Information Science and Technology (ARIST), 25: 209-260, 1990.

Hua Y. (2006). “Techniques for improved LSI text retrieval”. Ph.D Thesis, Wayne State University, 2006.

Jensen F. V. (1996). “Introduction to Bayesian Networks”. Springer.

Kai Gao, Yongcheng Wang and Zhiqi Wang. (2004). “An efficient relevant evaluation model in information retrieval and its application”. In International Conference on Computer and Information Technology, volume 0, pages 845{850, Los Alamitos, CA, USA. doi: http: //doi.ieeecomputersociety.org/10.1109/CIT.2004.1357300

Kontosthathis A. and Pottenger W. M. (2006). “A Framework for Understanding Latent Semantic Indexing Performance”. Journal of Information Processing and Management, 42, pp. 56–73.

Mooers C. N., “The theory of digital handling of non-numerical information and its implications to machine economics”. in Association for Computing Machinery Conference, Rutger University, 1950.

Pearl, J. (1988). “Probabilistic Reasoning in Intelligent systems: Networks of plausible inference”. Morgan Kaufmann Publishers Inc.

Reggia J. A. and Sutton G. G. (1988). “Self-processing networks and their biomedical implications”. Proceedings of the IEEE,Volume: 76, Issue: 6, Jun 1988.

Salton, G. and McGill, M. J. (1983). “Introduction to Modern Information Retrieval”. McGraw Hill, New York.

Turtle H. and Croft W.B. (1991). "Evaluation of an inference network-based retrieval model". ACM Transactions on Information Systems (TOIS), Volume 9, Issue 3, Pages 187-222, July 1991.

Wong S. K. M., Ziarko W. and Wong C. N. P. (1985). "Generalized Vector Space Model in Information Retrieval". Proceedings of 8th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 18–25.