2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

A Survey in Traditional Information Retrieval Models

Hai Dong[1], Farookh Khadeer Hussain[2], Elizabeth Chang[3]

[1]Hai Dong, DEBII, Curtin University of Technology, Perth, Australia, e-mail: hai.dong@cbs.curtin.edu.au
[2,3]Farookh Khadeer Hussain and Elizabeth Chang, DEBII, Curtin University of Technology, Perth, Australia
e-mail: (farookh.hussain, elizabeth.chang)@cbs.curtin.edu.au

***Abstract*—As a matter of fact, many so-called semantic search algorithms are derived from the traditional index-term-based search models. In this paper, we survey the traditional information retrieval models by categorizing them into three main classes and eleven subclasses, and analyse their benefits and issues of them.**

***Index Terms*—index-term based information retrieval models, Boolean models, algebraic models, probabilistic models.**

I. Introduction

In many semantic search engines and methods, the utilized semantic search algorithms are derived from the traditional index-term-based information retrieval models. In fact, many so-called semantic search algorithms are the combination or the evolved version of the traditional information retrieval algorithms. In this paper, we survey the traditional information retrieval models and analyse benefits and issues of them.

II. Semantic Search Models

The traditional information retrieval models, as in Fig.1, can be primarily divided into the category of set theoretic models, algebraic models and probabilistic models. Set theoretic models have four main types – Boolean model, case-based reasoning model, fuzzy set model and extended Boolean model. Algebraic models include vector space model, generalized vector space model, latent semantic indexing model and neural network model. Probabilistic models involve probabilistic model, inference network model, and brief network model. In the following sections, we will analyse these models and survey their benefits and issues.

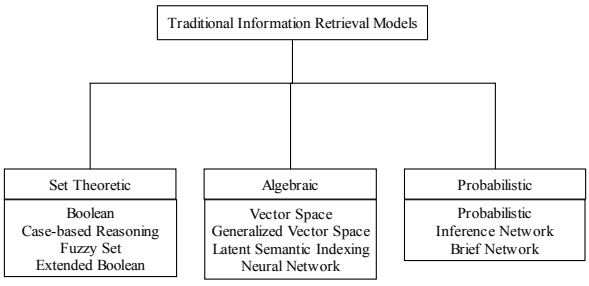

Fig.1 Traditional information retrieval models classification

III. Set Theoretic Models

A. Boolean model

Boolean algorithm is based on set theory and Boolean algebra. A set is a collection of abstract objects, where each object is the member of this set. Boolean algebra is a set of logical operations between two sets, such as conjunction, disjunction and complement [1].

In the Boolean model, whether an index term appears in a document or not determines the value of the weight between the index term and the document, which is a binary value. A query normally consists of several index terms connected by a set of logical operations, and it can be transferred to a conjunctive normal form which is composed of a number of conjunctive components. If any conjunctive component from a query has the counterpart in a document, the similarity of the document and the query is matched, and a value 1 is awarded; otherwise a value 0 is awarded [1].

B. Case-based reasoning (CBR) Model

CBR model is used to retrieve and reuse the existing problem solutions for emerging problems, which has four sub-processes as below [2]:

Retrieve: a new problem is matched with similar cases in database [2].

Reuse: if these cases are matched, the solutions to the retrieved cases are reused as the solutions of the emerging problem [2].

Revise: if the retrieved cases cannot completely match the problems, the solutions to the problem need to be revised [2].

Retain: the new case, incorporating both problems and solutions, is stored in database for further use [2].

The CBR matching algorithm is derived from the Boolean model [2]. Every feature extracted from incident reports is awarded an equal weight. Every feature in a new incident is compared with the corresponding feature in each of the other incidents. If the features match, a score of 1 is awarded. If the features do not match, a score of 0 is awarded. A similarity score is calculated as follows:

1. Finding the sum of the matching features [2]; and
2. Dividing this sum by the number of features contained in the incident [2].

Then a threshold is set up to determine whether the two incidents are matched or not [2].

2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

C. Fuzzy set model

Fuzzy set theory is to deal with the terms whose boundaries are not well defined. Each term in a query can be defined as a fuzzy set and has a degree of membership (between 0 and 1) with each document [17].

In the fuzzy set model, a thesaurus can be defined as a term-term correlation matrix in which rows and columns are related to the index terms [4]. The elements of the matrix are the correlation values between two terms. Moreover, this correlation matrix can be used to define a fuzzy set of documents related to each index term. In this fuzzy set, a document has a degree of membership with an index term, which is computed as an algebraic sum of all terms in the document, and if the algebraic sum value is beyond a threshold, the document belongs to the fuzzy set of the term. In order to match the semantic similarity between documents and a query, the query expression is converted to a set of conjunctive components. Then, each conjunctive component associates with a fuzzy set of documents and the union of the fuzzy sets are processed by Boolean operations. Finally, the membership value of each document in the processed fuzzy set is computed and ranked.

D. Extended Boolean model

Extended Boolean model is to extend the Boolean model with the function of matching and term weighting, which is a hybrid model to combine the Boolean model and the vector space model [9].

By weighting the association between a document and a query term by tf-idf algorithm, the similarity between a conjunctive query or a disjunctive query and a document can be calculated [1].

IV. ALGEBRAIC MODELS

A. Vector space model (VSM)

In VSM, each document is represented as a vector, and each dimension of the vector corresponds to a term in indexed terms [8] [10]. If a term appears in a document, a weight is assigned to a corresponding dimension in the vector. Similarly a query also can be seen as a vector with corresponding indexed terms. Thus, the relevance between a query and a document can be calculated as the cosine of the angle between the two vectors. A threshold is set up to determine the similarity between the terms from a document and a query, which enables a document to become partially similar to a query.

The weight can be calculated in terms of term frequency-inverse document frequency (tf-idf) weight [11]. Term frequency is the raw frequency of a term in a document, which is often normalized to prevent a bias towards longer documents. Inverse document frequency is the inverse of the frequency of a term among the documents in a system.

B. Generalized vector space model (GVSM)

Compared with VSM, index terms in GVSM are independent. Independency of index terms in GVSM means that the set of vectors is linearly independent and forms a subspace of interest [16].

In the GVSM, two vectors may be not orthogonal, but they are composed of smaller components which are derived from the particular collection. If the weights of association between index terms and documents are all binary, all possible patterns of term co-occurrence can be represented by a set of $2t$ minterms. The GVSM is to introduce a set of pairwise orthogonal vectors associated with the set of minterms and to adopt the set of vectors as the basis for the subspace of interest. To determine term vectors associated with a term, the vectors for all minterms in which the term is in state 1 are summed up and normalized. For each minterm vector, a correlation factor is defined by summing up the weight associated with the term and each document whose term occurrence pattern coincides exactly with that of the minterm. A minterm is of interest only if there is at least one document in the collection which matches its term occurrence pattern. In the GVSM, these representations can be directly translated to the space of minterm vectors. The resultant document vectors and query vectors are then used to compute the ranking through a standard cosine similarity function [1].

C. Latent semantic indexing (LSI) model

LSI is to map each document and query vector into a lower dimensional space which is associated with concepts [3]. This is accomplished by mapping the index terms vectors into this lower dimensional space. The LSI proposes to decompose a term-document association matrix in three components using singular value decomposition. The first one is the matrix of eigenvectors derived from the term-to-term correlation matrix; the second one is the matrix of eigenvectors derived from the transpose of the document-to-document matrix; the third one is a $r \times r$ diagonal matrix of singular values where r is the minimum between the row and the column of the original matrix, and the rank of the term-document association matrix. Consider now that only s largest singular values of the third matrix are kept, along with their corresponding columns in the first and the third matrix while the rest singular values are deleted. The resultant matrix is the matrix of rank s, which is closest to the original matrix in the least square sense. The relationship between two documents in the reduced space of dimensionality s, can be obtained from the multiplication of the resultant matrix and its transpose. To rank documents with regards to a query, the query is modelled as a pseudo-document in the original term-document matrix. Assume the query is modelled as the document with number 0. Then the first row in the multiplication of the resultant matrix and its transpose provides the ranks of all documents with respect to this query.

D. Neural network model

The neural network model for IR includes three layers – the query terms, the document terms and the documents (Fig.2) [1] [14]. Firstly, the query term nodes send signals to the documents' term nodes, then the documents term nodes generate signals to the documents nodes. The docu-

2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

ments nodes might generate new signals which are directed to the documents' term nodes, then the documents' term nodes may generate these new signals to other documents nodes and repeat this process. The signals become weaker at each iteration cycle, and the spread activation process eventually halts.

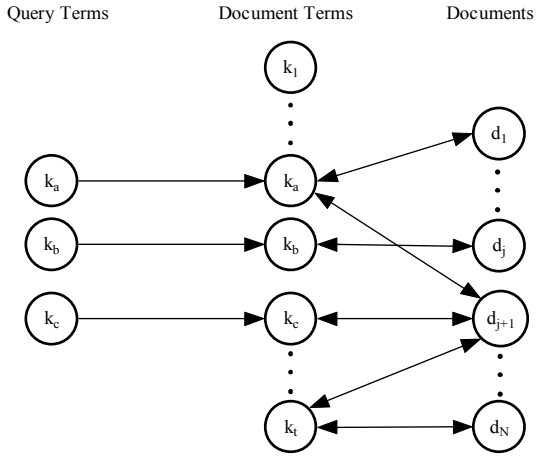

Fig.2 A neural network model for IR (adopted from [1])

The query term nodes are assigned an initial activation value as 1 [1]. The query term nodes then send signals to the documents term nodes which are attenuated by normalized query term weight. For a vector-based ranking, the normalized query term weight is derived from the weights defined by the vector model, where the normalization is done using the norm of query vector. Once the signals reach the document term nodes, these might send new signals to documents nodes. These signals are attenuated by normalized document term weight derived from the weights defined for the vector model where the normalization is done using the norm of query vector. After the first round of signal propagation, the activation level of the document node associated to the document is determined by the cosine algorithm provided by the VSM. A minimum activation threshold value is set up to improve the retrieval performance.

V. PROBABILISTIC MODELS

A. Probabilistic model

The core principle of a probabilistic model is to improve the probabilistic description of the ideal answer set of documents relating to a query by a series of interactions [6]. The process of a query is as follows:

Firstly, the user takes a look at the retrieved documents and initially guesses which one is related and which one is not, which groups a preliminary probabilistic description of the ideal answer set. Then the system can use this information to refine the description of the ideal answer set. By repeating this process many times, it is expected such a description will evolve and become closer to the real description of the ideal answer set [1].

The assumption, as the fundamental of probabilistic model, is as follows:

Given a user a query and a document in a collection, the probabilistic model tries to estimate the probability that the user will find the interesting document. The model assumes that this probability of relevance only depends on the query and the document representations only. Furthermore, the model assumes that there is a subset of all documents which the user prefers as the answer set for the query. Such an ideal answer set should maximize the overall probability of relevance to the user. Documents in the ideal set are predicted to be relevant to the query, and documents out of this set are predicted to be non-relevant [1].

B. Bayesian network

A Bayesian network is a directed acyclic graphic (DAG) model in which nodes represent a set of random variables and the arcs represent the conditional independencies between these variables (Fig.3) [1] [5]. The parents of a node are those judged to be the direct cause of the node, with a link directed from each parent node to the child node. The roots of the network are the nodes without parents. The Bayesian network can be used to improve the performance of probabilistic queries as it provides a complete formula for combining district sources of evidence in support of the rank for a given document. In the following, two main models based on the Bayesian model are introduced, which are the inference network model and the belief network model.

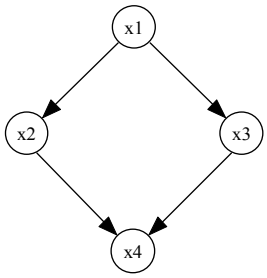

Fig.3 An example of Bayesian network (adopted from [1])

C. Inference network model (INM)

INM uses an epistemological view that "*interprets probability as a degree of belief whose specification might be devoid of statistical experimentation*" to deal with information retrieval issues [12] [13]. It associates random variables with the index terms, queries and documents. A random variable associated with a document means the event of observing the document by building a belief on associating with its index terms. Thus, the observation of a document causes an increased belief in the variables associated with its index terms.

Index term variables, query variables and document variables are represented as nodes in the network (Fig.4) [1]. Arcs directed from a document node to its term nodes represent that the observation of the document yields improved belief in its term nodes. The query variables means that the information request specified in these queries has been met. Arcs directed from index term nodes to query nodes means that the beliefs in the index term nodes associate with the query terms. The figure below illustrates the basic INM. An assumption in the network is that all random variables are binary.

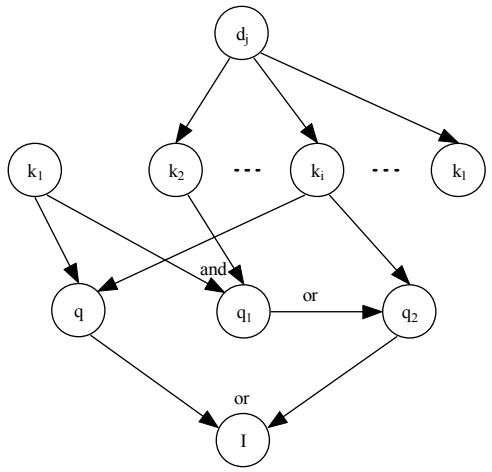

Fig.4 Basic inference network model (adopted from [1])

The ranking of documents with respect to a query is a measure of evidence supporting the observation of the documents towards the query. The ranking is obtained by basic conditioning and the application of Bayes' rule. The instantiation of a document node separates its children index term nodes, making them mutually independent. Thus, the degree of belief asserted to each index term node can be computed separately. As the documents nodes are the root nodes in the INM, they received a prior probability distribution chosen by users. The prior probability reflects the probability associated to the event of observing a given document. Since there is no prior preference for any documents in particular, a uniform prior probability distribution algorithm is adopted [1].

The INM covers a wide range of ranking strategies, including Boolean model and tf-idf algorithm [1].

For the Boolean model, the prior probabilities are all set to 1/N (N is the number of documents in the system) because the model does not make distinction on documents. When a document is observed, only the nodes associated with its index terms are active. The calculated beliefs in the index term nodes are used to compute the evidential support to a user query by considering that one of the conjunctive components of the user query must be exactly matched by the set of active terms [1].

For tf-idf algorithm, the prior probabilities are adopted to reflect the prior knowledge of the importance of document normalization. Normalized tf factors are taken into account through the beliefs asserted upon the index term nodes. Normalized idf factors are taken into account though the impact of index term nodes on the query nodes [1].

D. Brief network model (BNM)

BNM adopts a clearly defined sample space and thus it yields a sightly different network topology which provides a separation between the document and the query portions of the network [7].

The probability space is defined by a universe of discourse containing a set of index terms and all documents are indexed by the index terms [15]. Each index term is viewed as an elementary concept and the set is viewed as a concept space. A concept that is a subset of the set might represent a document in the collection or a user query. In the BNM, set relationships are specified by using random variables where 1 indicates that an index term is a member of a concept or a set represented by a vector associated with the set. A document in the collection can be represented as a concept composed of the terms which are used to index the document. A user query is represented as a concept composed of the terms which are used to index the query.

In the BNM, a user query is modelled as a network node associated to a binary random variable which is also referred to the document (Fig.5). This variable is set to 1 when the query completely covers a concept space. Thus, when the probability of a query is assessed, the degree of coverage of the concept space by the query is computed. Similarly, a document is modelled as a network node associated to a binary random variable which is also referred to the document. This variable is set to 1 when a document completely covers the concept space. When the probability of a document is assessed, the degree of coverage of the concept space by the document is computed [1].

Thus, the user query and the documents in the collection are modelled as subsets of index terms. Each of these subsets is interpreted as a concept embedded in the concept space which works a common sample space. Furthermore, user queries and documents are modelled identically. As the BNM in Fig.5, a query is modelled as a binary random variable pointed to by the index term nodes which compose the query concept. Documents are treated analogously to user queries. Thus, in contrast to the INM, a document node is pointed to the index term nodes which compose the document [1].

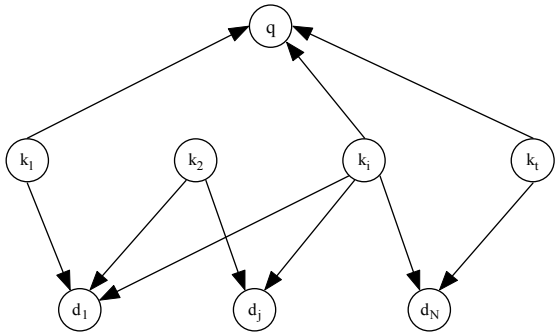

Fig.5 Basic belief network model (adopted from [1])

The ranking of a document relative to a given query is interpreted as a concept matching relationship and reflects the degree of coverage of the concept provided by the query concept [1].

VI. BENEFITS AND ISSUES OF SEARCH MODELS

In the following sections, the benefits and issues of these information retrieval models are summarized.

A. Benefits and issues of set theoretic models

The principle working in the Boolean methodology is simple, which saves time and cost of computing. However, the Boolean model's retrieval strategy is based on a binary criterion, which does not take into account fuzzy matches and semantic matches. Users also find difficulties in expressing their complex query requirement by means of Boolean operations. Finally there is no ranking method pro-

vided.

The CBR algorithm is simple and easy to implement. The returned similarity value is not a binary value but a decimal fraction between 0 and 1, which can be utilized in the ranking. However, the weight of features should be different according to their importance to belonged incidents. Furthermore, the algorithm does not take into account semantic matches.

Instead of directly using thesaurus to find the relevant terms, in other words, assigning same weight to each query terms and their synonyms, the fuzzy set theory is to compute the degree of semantic relevance between two terms, which is more efficient in improving precision. By using thesaurus, the documents that do not have index terms appearing can be retrieved. However, because the cost associated with computing the relevance between two terms counts on the number of occurrence of the terms in all documents, implementing the fuzzy set theory in large-scale databases is costly.

The extended Boolean model is an improvement to the drawback of the traditional Boolean model that does not have the ability to show the extent of relevance between query terms and index terms. It can be used in ranking of retrieved results, which can assist users to handle recall and precision. However, to extract and to maintain the index terms from dynamic sources in databases is costly on time. Moreover, the cost of computing the relevance between terms is vast.

B. Benefit and issues of algebraic models

In the VSM, the tf-idf term weighting scheme improves retrieval performance. Its partially matching strategy can improve the recall of retrieved results. The cosine algorithm is able to be used to rank and index the retrieved results. Nevertheless, the dynamical document bases make index terms difficult to maintain. In addition, the dependency of index terms is a prerequisite for VSM. Due to the locality of many term dependencies, the indiscriminate application to all documents in the collection might hurt the overall performance.

The GVSM can nicely represent the dependencies among index terms, which is the drawback of the VSM. However, the GVSM does not have a clear progress in practical performance, due to the fact hat the incorporation of term dependencies does not yield effective improvement with general collections. In addition, it is more complex and computationally more expensive than the VSM.

LSI forms an efficient indexing scheme for the documents in the collection, and it provides for elimination of noise and removal of redundancy. Nevertheless, currently the LSI has not been validated on the large scale document retrieval system.

The neutral network model provides an alternative searching paradigm. It also allows querying documents unrelated to the query terms, which is an attractive function, whereas this model has not been tested extensively with large document collections.

C. Benefits and issues of probabilistic model

In the probabilistic model, documents are ranked in decreasing order of their probability of being relevant. However, in the implementation of probabilistic model, users could make mistakes in initially guessing the initial separation of documents into relevant and non-relevant sets. Additionally, the method does not take into account the frequency of index terms in a document, which cannot weight the importance of different index terms towards a document. Finally the assumption that all index terms are independent is not feasible in practice.

The INM allows top retrieval performance to be accomplished with general collection. However, the computational cost of INM is as complex as the cost of computing a vectorial ranking. Moreover, it takes a purely epistemological view of IR problems, which is difficult to grasp.

The BNM is more general than INM which is only used in inquiry systems. It is based on the set theoretic view of the IR ranking problem and adopts a clearly defined sample space, which is easier to grasp than the INM. In addition, it provides a separation between the document space and the query space which simplifies the modelling task. Next, it facilitates the modelling of additional evidential sources, such as past queries and past relevant information. Finally it is able to reproduce any ranking strategy generated by the INM. The only drawback is that the computational cost of the BNM is as complex as the cost of computing a vectorial ranking.

VII. CONCLUSION

In this paper we have reviewed the existing information retrieval models. In our literature review, we classify the traditional information retrieval models into three main classes and eleven subclasses.

By analysing the issues, it is not difficult to conclude the primary issues in set theoretic models are the cost of computing (mainly in the fuzzy set model and the extended Boolean model), and lack of semantics (mainly in the Boolean model and the CBR model); the general issues in algebraic models are maintaining difficulties (mainly in the VSM), computational cost (mainly in the GVSM) and lack of validation in practice (mainly in the LSI and the neural network); the principle issues in probabilistic models are lack of practicality (mainly in the probabilistic model and the INM) and computing cost (mainly in the INM and BNM).

VIII. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Harlow: ACM Press, 1999.

[2] D. C. J. Carthy, A. Drummond, J. Dunnion, and J. Sheppard, "The use of data mining in the design and implementation of an incident report retrieval system," in *Systems and Information Engineering Design Symposium*, Charlottesville, 2003, pp. 13-18.

[3] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model for latent semantic structure," in *11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Grenoble, 1988, pp. 465-480.

[4] Y. Ogawa, T. Morita, and K. Kobayashi, "A fuzzy document retrieval system using the keyword connection matrix and a learning method," *Fuzzy Sets and Systems,* vol. 39, 1991.
[5] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* San Francisco: Morgan Kaufmann Publishers, Inc., 1988.
[6] B. A. Ribeiro-Neto and R. Muntz, "A brief network model for IR," in *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, 1996, pp. 235-260.
[7] S. E. Robertson and K. S. Jones, "Relevance weighting for search terms," *Journal of American Society for Information Science,* vol. 27, pp. 129-146, 1976.
[8] [8] G. Salton, *The SMART - Experiments in Automatic Document Processing*. Englewood Cliffs: Prentice Hall Inc., 1971.
[9] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communication of the ACM,* vol. 26, pp. 1022-1036, 1983.
[10] G. Salton and M. E. Lesk, "Computer evaluation for indexing and text processing," *Journal of the ACM,* vol. 15, pp. 8-36, 1968.
[11] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval* New York: McGraw Hill Book Co., 1983.
[12] H. Turtle and W. B. Croft, "Inference networks for document retrieval," in *the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Brussels, 1990, pp. 1-24.
[13] H. Turtle and W. B. Croft, "Evaluation of an inference-based retrieval model," *ACM Transactions on Information Systems,* vol. 9, pp. 187-222, 1991.
[14] R. Wilkinson and P. Hingston, "Using the cosine measure in a neural network for document retrieval," in *the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Chicago, 1991, pp. 202-210.
[15] S. K. M. Wong and Y. Y. Yao, "On modelling information retrieval with probabilistic inference," *ACM Transactions on Information Systems,* vol. 13, pp. 39-68, 1995.
[16] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, "Generalized vector space model in information retrieval," in *the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, 1985, pp. 18-25.
[17] L. A. Zadeh, "Fuzzy sets," in *Readings in Fuzzy Sets for Intelligent Systems*, D. Dubois, H. Prade, and R. R. Yager, Eds. San Francisco: Morgan Kaufmann, 1993.