Adding Term Weight into Boolean Query and Ranking Facility to Improve the Boolean Retrieval Model

Jiayi Wu
University of Windsor

There are two major shortcomings of the Boolean Retrieval Model that has attracted many researchers to improve the model since 1980s. These two problems are: inability to accommodate term weight in boolean queries and being unable to rank in Boolean Retrieval Model. Approaches to improve these two disadvantages can help us to achieve better precision and recall results for information retrieval. Various researchers have proposed approaches to solve the two problems. This survey reviews research which tries to clarify: (1). how well does the traditional boolean query work; (2). three different methods which can add term weight into boolean query; (3). the method of adding ranking facility to Boolean Model; (4). how well does the new extended model work and what are the evaluation results.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Information retrieval, term weight, boolean query, ranking facility

Contents

1. INTRODUCTION

As an essential model in information retrieval, boolean retrieval systems have been most widely used in different commercially available IR systems in terms of the simple query structure and effective results. This survey concerns research which attempts to give solutions to two major disadvantages of the boolean retrieval model. The two major disadvantages are the term weight problem and ranking facility problem. After adding the term weights into the boolean query and giving the ranking facility in boolean retrieval systems, this model will be more precise to retrieval information than before. The high precision means that the boolean retrieval systems will retrieve the relevant documents more accurately and this is one of the most basic goals of information retrieval.

Relevant research papers were found by searching Google Scholar with the keywords: “term weight”, “boolean query”, “ranking facility” and “boolean model”. Various keywords and author names were used to search the ACM publication library, IEEE, LNCS and other sources.

Fourteen journal papers, six conference papers, which are closely related to this survey were identified. They are listed in the bibliography.

Papers that were published in conference proceeding are Salton and Wu [1980], Wong et al. [1986], Lee et al. [1993], Pasi [1999], Latiri et al. [2003], Patro and Malhotra [2005]. Papers that were published in journals are Bookstein [1980], Kantor [1981], Buell [1982], Kraft and Buell [1983], Bordogna et al. [1991], Bordogna and Pasi [1993], Kraft et al [1994], Lee [1995], Monz [2007], Patro, Malhotra and Johnson [2007], Wu et al. [2008], Lioma and Blanco [2009], Pohl et al. [2012], Vignesh and Sivakumar [2013].

Ten papers were chosen as the basis of this survey. The papers and the reasons for choosing them are given below: Papers (Bookstein [1980], Kraft and Buell [1983] and Bordogna et al. [1991]) was chosen to identify fuzzy set approach of adding term weight into boolean query; Paper (Wong et al. [1986]) was chosen to identify the third method of adding term weight into boolean query; Paper (Lee et al. [1993]) was chosen to evaluate the new approach of adding ranking facility into boolean model; Papers (Lee [1995] and Pohl et al. [2012]) was chosen to identify the approach of adding ranking facility into boolean model; Paper (Pasi [1999]) was chosen to identify one of the approaches of adding term weight into boolean query; Paper (Patro and Malhotra [2005]) was chosen to identify the evaluation of the traditional boolean query; Paper (Monz [2007]) was chosen to analysis the query term weight.

The remainder of this survey is structured as follows: Section two contains reviews of seven papers (Bookstein [1980], Kraft and Buell [1983], Wong et al. [1986], Bordogna et al. [1991], Pasi [1999], Patro and Malhotra [2005], Monz [2007]). These papers are all concerned with analysis approaches which adding term weight into boolean query. Section three contains reviews of three papers (Lee et al. [1993], Lee [1995], Pohl et al. [2012]). These papers are all concerned with analysis approach of adding ranking facility into boolean model.

From all ten selected papers it is observed most of the authors used precision and recall to evaluate the effectiveness of their experiments and some of the authors used the average precision.

2. APPROACHES WHICH ON ADD TERM WEIGHT INTO BOOLEAN QUERY

Research papers that are presented in this section describe the importance of the term weight and give different approached which will apply term weight on boolean query. It is observed that researchers have used three different approaches. The most widely used approach was described by Bookstein [1980], Kraft and Buell [1983] and Bordogna et al. [1991]. There also another two approach described by Wong et al [1986] and Pasi [1999].

2.1 Fuzzy set method

Bookstein [1980] proposes the fuzzy request which is an approach to weight boolean search query. The author states that using boolean expressions permits one to represent accurately the logical relationships among concepts involved in an information need, but it has some loss in flexibility. When a user is able to express a concept in a boolean expression and its logical relationship to other concept, the user is not able to express how important that concept is to him relative to the other concepts represented in the query. The same situation will occur in documents indexing. Therefore, it is more desirable to have an approach in which one provides boolean queries with independently assigning weights to each term in the query to indicate how important that term is.

The author refers to previous work by Zadeh [1965].

The author states that Zadeh has worked on developing the concept of a fuzzy set to satisfy the need for a set that permits partial membership. However, it is hard to

determine whether a given document should be indexed by a specific term sometime and thereby be included in the set. So, for fuzzy sets only indicate the extent to which it is in a set.

The author states that the purpose of his paper is to propose a method for resulting information system merges some of the boolean and weighted systems being accomplished by relying on a generalization of the traditional algebra of sets and by defining a weighting scheme for requests that is consistent with this algebra. In this paper, the author gives a general idea of fuzzy set which is a new extended expression and the manipulations on fuzzy sets can be defined in terms of the membership functions: inclusion, union, intersection and complementation. Then, the author analyses how the fuzzy set works with the queries in which terms are weighted. The specific queries being analysis include four forms of boolean expressions: the single index terms, the terms are connecting by AND, the terms are connecting by OR and the terms are connecting by NOT. The author states that allow transforming queries into more convenient forms is one of characteristics of boolean retrieval systems. In this paper, the author gives some rules that permit one to change a fuzzy query into a different but equivalent one and some relationships follow immediately from the properties of fuzzy sets in general. The rules include: commutativity, associativity, distributivity, duality, idempotency, weight distributivity, involution and weight manipulation.

The author gives some example graphs of analysis of the four forms of boolean expression and use tables simply show the analysis results of them.

The author states that the weighted assignment of index terms can be modeled in terms of fuzzy sets so as to accomplish this goal in all four different boolean expressions.

Bookstein [1980] claims that it is possible to assign weights to the new expression when it can be transform into an equivalent Boolean expression.

This paper is cited by Bordogna et al. [1991].

Kraft and Buell [1983] applied fuzzy sets into the generalized Boolean retrieval systems. The authors state that the problem for traditional boolean retrieval model can be seen as decision theory problem. The basic assumptions of the model determine documents as relevant or totally non-relevant. It not possible to consider a document partially relevant and this situation may miss some important relevant documents. According to this problem, there are some new approaches has been generalized to allow for weights to be attached to individual terms, in either the document indexing or the query representation, or both.

The authors refer to previous work by Salton and Wu [1980].

The authors state that Salton has worked on vector model which includes the weight on index terms but queries are not boolean and are usually discrete. This method does not allow for boolean logic in the query structure that is lacking in general. In this model, differences are precisely what the fuzzy threshold approach emphasizes, while still allowing for boolean logic in the query.

The authors state that the concept of a fuzzy subset can be applied to the document retrieval situation. The membership function can describe terms with a weight in the interval [0, 1]. The traditional Boolean indexing has the weights with zero or one. If

weights are in the open interval (0, 1), then, the fuzzy indexing will be used. The probabilistic approach used in the fuzzy retrieval is concerned with estimating the probability of relevance of a given document to a query and it preserves the boolean lattice properties. The fuzzy subset retrieval allow user to generalize from single-term queries. So, boolean connectives are similar with the vector lattice. The generalization of traditional Boolean query processing is more complex than merely fuzzifying the indexing. The query representation can also be weighted. There are four types of generalizations: boolean indexing and boolean queries with non-Boolean retrieval status values, fuzzy indexing and boolean queries with retrieval status values computed using fuzzy subset rules, boolean indexing and fuzzy queries with retrieval status values and fuzzy indexing and fuzzy queries with the retrieval status value being calculated by some general function. The last generalization has problem with generating a "weight" for a term and document evaluation in terms of its relevance. A new an alternative function for document relevance evaluation has been mentioned.

The authors use a threshold approach, rather than a weight. This can solve the last generalization problem. A new function form conducted and several criteria which specify necessary conditions for a proper document evaluation mechanism have been mentioned.

The function form implies that one is given some partial credit (F/a) for the membership function's coming close but not exceeding the threshold. This credit is weighted by an increasing function of the threshold. This implies that as the threshold increases, a given percentage of partial is given more weight.

Kraft and Buell [1983] claim that fuzzy subset theory applies to document retrieval systems allowing non-boolean index weights to be attached to the document and non-Boolean weights or thresholds to be attached to the individual terms in the query representation. This is this is a generalization of document and query representation and processing.

This paper is cited by Bordogna et al. [1993]

Bordogna et al. [1991] proposed that query term weights as constrains in the fuzzy information retrieval. The authors state that the boolean model is widely used in many traditional systems and it suitable for a flexible query formulation; however, it also has some disadvantages. The boolean model retrieves information items only into two classes: relevant and irrelevant. Therefore the model does not allows the ranking property of documents in descending order of estimated by a query and does not provide a method to solve with the imprecision in the query formulation.

The author refers to previous work by Bookstein [1980], Kantor [1981], Buell [1982], Kraft and Buell [1983].

The authors state that Bookstein’s work on the fuzzy set model as the first kind of semantics defines query weights as measures of the relative relevance of each term with respect to other terms. This model defines the relevance semantics in all aspects but the AND operator is associated with the lowest weighted term. This situation generates contradiction with the semantics of relevance. The authors state that Kantor’s works is another difficulty that the notion of complementation when using relevance weights in the fuzzy set context. The authors state that Buell’s worked on another approach with

threshold semantics for query weights also have a problem: the meaning of threshold weights is not clear when apply to boolean expressions rather than to single terms and this is a problem of semantics.

The authors present an extended boolean model formally described by means of the fuzzy set theory and the problem of query weighting in the existing models. In the author's model, a query term weight has the meaning of an ideal document-term relation value which be considered as a constraint on the stored document representations. So, the system will retrieve first documents whose index term weight is close to w (where w is a weight to a term t in a query). This interpretation of query term weight as clear requirements of ideal index term weights permits interpretation of a query as description of one or more ideal documents for the user. The Retrieval Status Value (RSV) is expressed as the degree to which all constrains has been satisfied by each document representation in stored collection. The authors use fuzzy set theory to define the generalization of the boolean model and the constraint is expressed by the formalism of fuzzy restrictions.

RSV evaluation mechanism is defined and analyzed with respect to the Cater and Kraft wish-list. The author use a function E* to evaluate the matching of a query against a collection of documents.

The authors state that function E* satisfies these criteria for an RSV mechanism: separability, Boolean restriction, Self consistency, Term similarity, Weights of zero, Query weight volume, Binary Boolean operations and Unary Boolean operation.

The authors claim that an analytical approach to the interpretation of weighted boolean query has been presented in their paper. A query becomes a means of describing classes of ideal documents and expressing relativity criteria in order to distinguish query term weights from query weights.

This paper is cited by Kraft et al [1994].

2.2 On extending the vector space model for boolean queries

Wong et al. [1986] introduced the idea of extending the vector space model applies for boolean queries. The authors state that the boolean retrieval systems do not apply the incorporating term correlations into the retrieval process. In another words, the weighted queries and documents is a problem for Boolean retrieval systems.

The authors refer to previous work by Buell [1981].

The authors state that Buell's work on the strict Boolean retrieval systems has the problem that there is no provision for weights of importance to the terms both in the queries and documents. The representation is binary that lacks various index terms and the output is not ranked. In most cases, the AND connectives tend to be too restrictive.

The authors introduce a new information retrieval model, named Generalized Vector Space Model (GVSM). The new model solves the weights problem in the boolean model and the queries used in this model are seemed as an extended boolean expressions. In GVSM, a query is simply defined as a weighted vector sum of term vectors. However, this form of query does not help the user to explain clearly structure as can be done in boolean systems. Therefore, the most common language used to express query structure

involves boolean logic and the query is considered to be a list of index term and weight pairs. This is the first important variation queries that are called the basic GVSM. The other one is the called unified GSVM which involves a scheme for expressing weighted Boolean queries as vectors. Queries are specified as a weighted boolean expression in which are connected by AND, OR and NOT operators.

The authors compare the unified model with the p-norm model which was applied for extended boolean retrieval model. The authors use MEDLARS and CISI collections for experimental evaluation, because other collections used for information retrieval do not provide boolean queries. The standard recall and precision measures are used for comparing the performance of different models.

After comparing these models, the authors find that both models handle weighted boolean queries, both reduce to VSM and strict Boolean retrieval models under certain conditions, p-norm model involves the parameter p, which has to be experimentally determined and the extended query language satisfies more algebraic properties under unified GVSM. The unified GVSM is closer to the strict Boolean model than it is to VSM which is means the roles of Boolean operators is rather strictly retained, however, p-norm model achieves more softening of the operators.

Wong et al. [1986] claim that it would be advantageous to provide a prescription to handle boolean queries in the GVSM environment. The important part of basic GVSM is generalizing VSM to incorporate term correlations and document representation reduces to the vector sum of terms when terms are assumed to be orthogonal. The unified GVSM reduces to the strict Boolean retrieval model when each document is represented by its dominant atomic vector.

This paper is cited by Wu et al. [2008]

2.3 A logical formulation of weighted boolean models

Pasi [1999] proposes a logical formulation of weighted boolean models. The author states that in order to model IR in the logical framework there is a need for a more general formal discipline. It is necessary to analyse the role of logic in IR by defining a model of information retrieval based on modal logics, which provides a general framework to define pre-existing IR models. The query evaluation process of the boolean model and of weighted Boolean models will be exploited by analyzing the role of logic as a formal basis. The analysis will give better understanding of some query weight semantics.

The author does not refer to any of the papers in the bibliography as related work.

The author gives a formulation of the boolean model that expresses the evaluation structure of the boolean query. Fuzzy implications can be employed to generalize the logical interpretation of the Boolean model and it is necessary to describe the extension of the boolean model. An extended boolean model gives a weighted indexing function that evaluating a query term and the index term weight and it is interpreted as the degree of relevance of document with respect to query term. To enrich the expressiveness of the boolean query language, numeric query weights have been introduced as an extension of the basic selection criteria, which become then pairs term-weight.

The author uses QT (t, q) as the terms appearing in a given query q and IT (t, d) as the index terms belonging to the representation of document d. So, the logical interpretation of a Boolean query is the formal expression of the constraint imposed by a query term: QT(t, q) → IT (t, d). "The expression representing the query evaluation structure of the Boolean query q = (t_1 AND t_2 OR (NOT t_3)) is the following: (QT (t_1, q) → IT(t_1, d) ∧ QT(t_2, q) → IT(t_2, d)) ∨(¬ QT(t_3, q) → IT(t_3, d))." There is an extension of the logical interpretation of the Boolean model to weighted Boolean models. The IMP and QW is the importance of the index terms in document and in query representations. "A query of the type q = < t_1, w_1 > AND < t_2, w_2 > OR < t_3, w_3 >, has an evaluation which is formally expressed by the following logical expression: ((QW(t_1, q) → IMP(t_1, d)) ∧ (QW(t_2, q) → IMP(t_2, d))) ∨ (QW(t_3, q) → IMP(t_3, d))."

In the interpretation, the weighted query is a part of the formulation. The constant symbols are terms in the formulation and the logical connectives correspond to the Boolean connectives. The choice of the implication operator is important in the formalization, as it is strictly connected with the semantics of the query term-weight.

Pasi [1999] claims that the approach is based on the following considerations: terms are the most essential elements which evaluate the relevance in a query; the natural logical interpretation of the boolean model is important and the degree of relevance related to the truth of the given interpretations. The author also claims that the most important part of this approach is to make the bottom-up structure of the query evaluation procedure clear and the implication connective is employed to express limitation controlled by a query term on the document representations.

This paper is cited by Latiri et al. [2003].

2.4 Evaluation of traditional boolean query

Patro and Malhotra [2005] estimated the success from characteristics of the boolean web search query. The authors state that whether the query will successfully satisfy the users' requirements are unknown. Similarly, whether the Boolean query is efficient or not in finding the location of the best documents is unknown.

The authors do not refer to any of the papers in the bibliography as related work.

The authors state that this paper relates characteristics of the search results which are meet the user's requirement. A program which compares the performance of humans and the search queries synthesized performed better than humans and gives an objective judgment of the search queries.

The author states that use the precision and recall to measure the quality of queries is the common method. When search query, use the number of relevant documents first returned 20 links as the precision measurement. The authors use the same range as precision to identify recall. The authors collect data by using volunteers. Volunteers are students have different topic area and some sample relevant documents. After they use the same search engine to search their topic, they revised their query. Then compare the volunteer's query and the synthesized query. The authors use figures to show the relationship between recall and precision of the volunteer queries; the relationship between recall and precision after the queries' quality changed; the precision as function

of terms in query and number of attempts to improve query and relationship between number of terms and precision.

A good query returns higher values for precision and good recall which can be used as evaluation results of the method applied to the boolean model. A good predictor of a successful Boolean web search query is four terms and above average recall.

Patro and Malhotra [2005] claim that they found the characteristics of good queries that may give good sample of the query successfully achieved by users and their requirements.

This paper is cited by Patro, Malhotra and Johnson [2007].

2.5 Learning for query term weighting

Monz [2007] proposed a model tree learning for term weight. The author states that the procedure to find an answer may be stuck when the retrieval system failed to find relevant information. Therefore, the effectiveness of retrieval in a question answering system is important and the formulation of the queries has dominant influence in those retrieval components, especially in Boolean retrieval. Using the wrong query terms may lead a processing retrieval into a wrong way.

The authors refer to previous work by Brill et al. [2002].

The author states that Brill worked on retrieving the documents which match the query string or boolean query but do not solve the weight term problem.

One of the author's ideas is to use different query structures to discriminate terms which have positive or negative influence in the effectiveness of question retrieved. By ranking different well performing query variants and calculating the term weight (presence and absence weights) in the higher ranked query variants, the positive or negative gain can be found. The other approach used to solve the problem is derived from the first approach: the same term can be positive or negative gain in different queries but cannot give a term weight which did not occur. By using the representative features which have the term's information in the question instead the actual terms. In addition, use of a decision tree machine learning method to learn the query term's usefulness for query formulation and then give the weight to the query term.

The author first used TREC-9, TREC-10, and TREC-11 data sets to compare different ways of choosing terms from the questions and try to identify the optimal query term. And then use different enhanced calculation method to analysis term weight.

The author states that the query structure has a dominant influence in process retrieval system from the comparison experiments' results and better query structure can be performance better. From all the experiments, there are two factors affecting the query term weight: the term frequency in a documents and the retrieved frequency.

Monz [2007] claims that his work demonstrates the importance to learn the query term weight in query structure as well as the term selection will have dominant influence in the performance of the retrieval components that confirmed by decision tree. The author is also confident about using the approach in the paper will benefit the question answering systems.

This paper is cited by Lioma and Blanco [2009].

2.6 Summary

Year	Title	Authors	Major contribution
1980	Fuzzy Requests: An Approach to Weighted Boolean Searches.	Bookstein.	A general idea of fuzzy set which is a new extended expression and the manipulations on fuzzy sets can be defined in terms of the membership functions.
1983	Fuzzy sets and generalized Boolean retrieval systems.	Kraft and Buell	Describe that concept of a fuzzy subset can be applied to the document retrieval situation.
1986	On extending the vector space model for Boolean query processing.	Wong et al.	A new information retrieval model, named Generalized Vector Space Model (GVSM).
1991	Query term weights as constraints in fuzzy information retrieval.	Bordogna et al.	Define the generalization of the boolean model use fuzzy set theory and the constraint is expressed by the formalism of fuzzy restrictions.
1999	A logical formulation of the Boolean model and of weighted Boolean models.	Pasi, G.	A new formulation of the boolean model that expresses the evaluation structure of the boolean query.
2005	Characteristics of the Boolean Web search query: Estimating success from characteristics.	Patro and Malhotra.	A program which compare the performance of human and the search queries synthesized performed better than human and it gives an objective judgment of the search queries.
2007	Model Tree Learning for Query Term Weighting in Question Answering.	Monz.	Use a decision tree machine learning method to learn the query term's usefulness for query formulation and then give the weight to the query term.

3. RANKING FACILITY TO THE BOOLEAN MODEL

As mentioned earlier the ranking facility does not apply in the traditional boolean model that will generate some problems of the retrieval results. An Extended Boolean Model is described in this section and the new model contains the ranking facility which improves the tradition model with better retrieval results.

3.1 Evaluation of Extended Boolean Operators

Lee et al. [1993] evaluated the boolean operators in the Extended Boolean Retrieval Framework. The authors state that the boolean retrieval systems have been the commonly used information retrieval system because of the efficient retrieval results and easy query structure. But the ranking is not supported and similarity coefficients cannot be calculated between queries and documents in traditional boolean retrieval systems. The fuzzy set model and the extended boolean model have been suggested providing the ranking function to the traditional boolean system and they are logical extensions of Boolean model because they reduce to boolean model when document term weights are restricted to zero or one. But still two problems exist: incorrect ranked output in some case and complex computation on boolean operators.

The authors refer to previous work by Bookstein, [1980] and Salton [1989].

The authors state that Bookstein's work on the fuzzy set model generates incorrectly ranked output in certain cases because the MIN and MAX operators have properties adverse to retrieval effectiveness. T-Operators in the fuzzy set theory have the Single Operand Dependency Problem and Negative Compensation Problem. Salton's work on the extended Boolean model has solve the problem of the former by apply the E_{AND} and E_{OR} operators; however, it suffers from the complex computation.

The authors give the T-operators and the corresponding operator graphs, and also, the averaging operators and corresponding operator graphs. Then using the graphs to describe that one pair of the Averaging operators of Fuzzy Sets Model and the operators of Extended Boolean Model overcomes the single operand dependency and negative compensation problems. These operators are defined as positively compensatory operators.

The authors use two different document collections covering items the ISI collection and the CACM collection. They compare the precision of each operator such as T-operator and average operator to find the effectiveness of them. The authors use the document term weights to evaluate rank documents in Extended Boolean Retrieval Framework. "The weight of document is normalized as follow: W_{ik} = (TF_{ik}/maximum TF in document i)* (IDF_k maximum/IDF in document i) where Inverse Document Frequency define as IDF and Term Frequency define as TF."

The authors state that positively compensatory operators provide higher retrieval effectiveness than the others. One pair of the Averaging operators of Fuzzy Sets Model achieves similar retrieval effectiveness and higher retrieval efficiency in comparison with the operators of Extended Boolean Model.

Lee et al. [1993] claim that the extended Boolean model has overcome the single operand dependency problem of the fuzzy set model by developing the operators for the evaluation of the AND and OR operations.

3.2 Effectiveness of Extended Boolean Model

Lee [1995] analyzed the extended Boolean models. The author states that each document is indexed with a set of keywords or terms, and each query contains terms connected with the Boolean operators AND, OR and NOT in the traditional Boolean model. However, the traditional model does not provide a document ranking function because it

cannot compute similarity coefficients between query and documents and this function reflects the relevance between query and documents that has the same objective with term weight.

The author refers to previous work by Sachs [1976], Radecki [1979] and Buell [1980].

The author states that MIN and MAX operators have been developed to support ranking function in the past for Boolean retrieval system. However, they do not correspond well with human behavior for the calculation of query-document similarities and lead the fuzzy set model to generate incorrect ranked output in some cases.

The author analyzes the behavioral aspects of different operators for AND and OR operations and the four important properties of retrieval effectiveness: single operand dependency, negative compensation, double operand dependency and unequal importance. The author describes them through examples and the four important could decrease retrieval effectiveness in some circumstance. The author also suggests that the properties of positive compensation retrieval and equal importance may help retrieval effectiveness. The author defines an operator class called n-ary soft Boolean operators that is suitable for achieving high retrieval effectiveness.

The author evaluates the effectiveness of sixteen different operators that can evaluate AND and OR operations in extended boolean models. The author uses a new large data collection to evaluate the effectiveness of the different operators in extended boolean models. The large data collection includes one of the TREC sub-collections which is called WSJD2. The author also uses document term weights to calculate document values and two famous weighting schemes have been used which are Fox-weights and INQUERY-weights.

The author states that the experiment's results suggest that the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem. The double operand dependency problem as well as the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem. INQUERY-weights give better retrieval effectiveness to the fuzzy set and Waller-Kraft models than Fox-weights. Network Boolean has the positively compensatory property in some operand values. The effectiveness of the p-norm model is slightly better than the vector space model for the well-formulated Boolean query.

Lee [1995] claims that this paper does not present optimal operators, but the properties being analyzed can be used as a high base line to approach optimal operators.

This paper is cited by Pohl et al. [2012].

Pohl et al. [2012] also analyzed extended Boolean models. Boolean queries have the disadvantage of being harder to formulate than ranked queries. They have the drawback of generating answer lists of unpredictable length and changes in the query that appear to be small might result in disproportionately large changes in the size of the result set. The queries of Extended Boolean Retrieval (EBR) models, such as the p-norm model, queries are slow to evaluate, because of their complex scoring functions and none of the computational optimizations available for ranked keyword retrieval have been applied to EBR.

The authors do not refer to any of the papers in the bibliography as related work.

The authors describe a scoring method for EBR models and adopts ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions. The authors also present the p-norm EBR model as an instance of such models and that performance gains can be attained that are similar to the ones available when evaluating ranked queries. Term-independent bounds are proposed in this paper, which complement the bounds obtained from max-score. It can be employed in the wand algorithm, also reducing the number of score evaluations.

The authors evaluated the efficiency of their methods on a large collection of biomedical literature using queries and results derived from real searches. Three query sets were used against this collection: 50 simple, short PUBMED queries consisting of a Boolean conjunction only, 50 structured queries containing both conjunctive and disjunctive operators at least once in each query sampled from the same query log, 15 complex queries. Properties of these queries are summarized. The authors counted the number of scored documents with scores below the entry threshold and above the entry threshold.

The authors state that the proposed scoring method is often being faster and it significantly reduces the number of candidate documents scored, postings processed, and execution times, for all query sets. Term-independent bounds method for short-circuiting candidate document scoring reduce the number of score calculations, especially on simpler queries and when combined with max-score. The query execution times of boolean execution will be faster, however, it must be remembered that the result of a Boolean query is of indeterminate size.

Pohl et al. [2012] claim that optimization techniques developed for ranked keyword retrieval can be modified for EBR and this leads to considerable speedups. Term-independent bounds provide added benefit when complex scoring functions are used and it will as a mean for short-circuit score calculations.

This paper is cited by Vignesh and Sivakumar [2013].

3.3 Summary

Year	Title	Authors	Major contribution
1993	On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework.	Lee et al.	Compare T-operators and averaging operators to find the positively compensatory operators.
1995	Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval.	Lee.	Analyzes the behavioural aspects of different operators and the four important properties of retrieval effectiveness.
2012	Efficient Extended Boolean Retrieval.	Pohl et al.	A scoring method for Extended Boolean Retrieval

			Model and adopts ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions.

4. CONCLUSIONS

Ten important papers	Ten important papers refered to others
Bookstein 1980	Bordogna et al. 1991
Kraft and Buell 1983	Salton and Wu 1980, Bordogna et al. 1993
Wong et al. 1986	Wu et al. 2008
Bordogna et al. 1991	Kantor 1981, Buell 1982 and Kraft et al. 1994
Lee et al. 1993	Pohl et al. 2012
Lee 1995	Pohl et al. 2012
Pasi 1999	Latiri et al. 2003
Patro and Malhotra 2005	Patro, Malhotra and Johnson 2007
Monz 2007	Lioma and Blanco 2009
Pohl et al. 2012	Vignesh and Sivakumar 2013

This survey identified 20 papers related to the topic of the survey and reviewed 10 of them in detail. In doing this with have made the following observations:

For the fuzzy set theory is a commonly used method to add term weight into boolean query and also be described as a fuzzy retrieval model which based on the boolean model. Many researchers such as Bookstein [1980], Kraft and Buell [1983] and Bordogna et al. [1991] have analysed this theory and attempt to improve the "Advanced Boolean Model". The fuzzy set model has been amplified and refined from the last century up to now. This model gives the term weight into the query that always used in the vector space model. Thus, Wong [1986] proposed a new method applying to boolean model by extend the vector space model. This theory gave a good combination the two traditional models. It not only improved and perfected the boolean model but also had the same effect on the vector space model.

The Extended Boolean Model the only one that applies the ranking facility based on boolean retrieval model. However, Lee [1995] proposed that different operators will have different behavioral aspects. The new operators in Extended Boolean Model can be defined as negatively operators and positively compensatory operators. A scoring method for Extended Boolean Retrieval Model also has been proposed and analyzed in recent years.

5. ACKNOWLEDGEMENT

During the completion of the survey, I have gained much support and instructions from many people. I would like to take this opportunity to thank them all.

First, I would like to thank Dr. Richard Frost. It is his continuous instructions, supports and encouragements that lead me to accomplish this survey. I really thank him for instructions and lectures.

Then, I would like to extend my thanks my supervisor Dr. Joan Morrissey. Thank to her for your advices and suggestions during the time I produced this survey.

Last, thanks for all the people who have given support to my survey. Thank you very much.

6. ANNOTATIONS

6.1 Bookstein 1980

Citation:

BOOKSTEIN, A. 1980. Fuzzy Requests: An Approach to Weighted Boolean Searches. *American Society for Information Science*. 31(4), 240.

Problem: The author states that using boolean expressions permits one to represent accurately the logical relationships among concepts involved in an information need, but it has some loss in flexibility. When a user is able to express a concept in a boolean expression and its logical relationship to other concept, the user is not able to express how important that concept is to him relative to the other concepts represented in the query. The same situation will occur in documents indexing. Therefore, it is more desirable to have an approach in which one provides boolean queries with independently assigning weights to each term in the query to indicate how important that term is.

Previous work: The author refers to previous work by Zadeh [1965].

Shortcomings of Previous work: The author states that Zadeh has worked on developing the concept of a fuzzy set to satisfy the need for a set that permits partial membership. However, it is hard to determine whether a given document should be indexed by a specific term sometime and thereby be included in the set. So, for fuzzy sets only indicate the extent to which it is in a set.

New Idea/Algorithm/Architecture: The author states that the purpose of his paper is to propose a method for resulting information system merges some of the boolean and weighted systems being accomplished by relying on a generalization of the traditional algebra of sets and by defining a weighting scheme for requests that is consistent with this algebra. In this paper, the author gives a general idea of fuzzy set which is a new extended expression and the manipulations on fuzzy sets can be defined in terms of the membership functions: inclusion, union, intersection and complementation. Then, the author analyses how the fuzzy set works with the queries in which terms are weighted. The specific queries being analysis include four forms of boolean expressions: the single index terms, the terms are connecting by AND, the terms are connecting by OR and the terms are connecting by NOT. The author states that allow transforming queries into more convenient forms is one of characteristics of boolean retrieval systems. In this paper, the author gives some rules that permit one to change a fuzzy query into a different but equivalent one and

some relationships follow immediately from the properties of fuzzy sets in general. The rules include: commutativity, associativity, distributivity, duality, idempotency, weight distributivity, involution and weight manipulation.

Experiments Conducted: The author gives some example graphs of analysis of the four forms of boolean expression and use tables simply show the analysis results of them.

Results: The author states that the weighted assignment of index terms can be modeled in terms of fuzzy sets so as to accomplish this goal in all four different boolean expressions.

Claims: The author claims that it is possible to assign weights to the new expression when it can be transform into an equivalent Boolean expression.

Citation by others: Bordogna et al. [1991].

6.2 Bordogna et al. 1991

Citation:

BORDOGNA, G., CARRARA, P. AND PASI, G. 1991. Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27(1), 15-26.

Problem: The authors state that the boolean model is widely used in many traditional systems and it suitable for a flexible query formulation; however, it also has some disadvantages. The boolean model retrieves information items only into two classes: relevant and irrelevant. Therefore the model does not allows the ranking property of documents in descending order of estimated by a query and does not provide a method to solve with the imprecision in the query formulation.

Previous work: The author refers to previous work by Bookstein [1980], Kantor [1981], Buell [1982], Kraft and Buell [1983].

Shortcomings of Previous work: The authors state that Bookstein's work on the fuzzy set model as the first kind of semantics defines query weights as measures of the relative relevance of each term with respect to other terms. This model defines the relevance semantics in all aspects but the AND operator is associated with the lowest weighted term. This situation generates contradiction with the semantics of relevance. The authors state that Kantor's works is another difficulty that the notion of complementation when using relevance weights in the fuzzy set context. The authors state that Buell's worked on another approach with threshold semantics for query weights also have a problem: the meaning of threshold weights is not clear when apply to boolean expressions rather than to single terms and this is a problem of semantics.

New Idea/Algorithm/Architecture: The authors present an extended boolean model formally described by means of the fuzzy set theory and the problem of query weighting in the existing models. In the author's model, a query term weight has the meaning of an ideal document-term relation value which be considered as a constraint on the stored document representations. So, the system will retrieve first documents whose index term weight is close to w (where w is a weight to a term t in a query). This interpretation of query term weight as clear requirements of ideal

index term weights permits interpretation of a query as description of one or more ideal documents for the user. The Retrieval Status Value (RSV) is expressed as the degree to which all constrains has been satisfied by each document representation in stored collection. The authors use fuzzy set theory to define the generalization of the boolean model and the constraint is expressed by the formalism of fuzzy restrictions.

Experiments Conducted: RSV evaluation mechanism is defined and analyzed with respect to the Cater and Kraft wish-list. The author use a function E* to evaluate the matching of a query against a collection of documents.

Results: The authors state that function E* satisfies these criteria for an RSV mechanism: separability, Boolean restriction, Self consistency, Term similarity, Weights of zero, Query weight volume, Binary Boolean operations and Unary Boolean operation.

Claims: The authors claim that an analytical approach to the interpretation of weighted boolean query has been presented in their paper. A query becomes a means of describing classes of ideal documents and expressing relativity criteria in order to distinguish query term weights from query weights.

Citation by others: Kraft et al. [1994].

6.3 Kraft and Buell 1983

Citation:

KRAFT, D. H. AND BUELL, D. A. 1983. Fuzzy sets and generalized Boolean retrieval systems. *International Journal of Man-Machine Studies*, 19(1), 45-56.

Problem: The authors state that the problem for traditional boolean retrieval model can be seen as decision theory problem. The basic assumptions of the model determine documents as relevant or totally non-relevant. It not possible to consider a document partially relevant and this situation may miss some important relevant documents. According to this problem, there are some new approaches has been generalized to allow for weights to be attached to individual terms, in either the document indexing or the query representation, or both.

Previous work: The authors refer to previous work by Salton and Wu [1980].

Shortcomings of Previous work: The authors state that Salton has worked on vector model which includes the weight on index terms but queries are not boolean and are usually discrete. This method does not allow for boolean logic in the query structure that is lacking in general. In this model, differences are precisely what the fuzzy threshold approach emphasizes, while still allowing for boolean logic in the query.

New Idea/Algorithm/Architecture: The authors state that the concept of a fuzzy subset can be applied to the document retrieval situation. The membership function can describe terms with a weight in the interval [0, 1]. The traditional Boolean indexing has the weights with zero or one. If weights are in the open interval (0, 1), then, the fuzzy indexing will be used. The probabilistic approach used in the fuzzy retrieval is concerned with estimating the probability of relevance of a given document to a query and it preserves the boolean lattice properties. The fuzzy subset

retrieval allow user to generalize from single-term queries. So, boolean connectives are similar with the vector lattice. The generalization of traditional Boolean query processing is more complex than merely fuzzifying the indexing. The query representation can also be weighted. There are four types of generalizations: boolean indexing and boolean queries with non-Boolean retrieval status values, fuzzy indexing and boolean queries with retrieval status values computed using fuzzy subset rules, boolean indexing and fuzzy queries with retrieval status values and fuzzy indexing and fuzzy queries with the retrieval status value being calculated by some general function. The last generalization has problem with generating a "weight" for a term and document evaluation in terms of its relevance. A new an alternative function for document relevance evaluation has been mentioned.

Experiments Conducted: The authors use a threshold approach, rather than a weight. This can solve the last generalization problem. A new function form conducted and several criteria which specify necessary conditions for a proper document evaluation mechanism have been mentioned.

Results: The function form implies that one is given some partial credit (F/a) for the membership function's coming close but not exceeding the threshold. This credit is weighted by an increasing function of the threshold. This implies that as the threshold increases, a given percentage of partial is given more weight.

Claims: The authors claim that fuzzy subset theory applies to document retrieval systems allowing non-boolean index weights to be attached to the document and non-Boolean weights or thresholds to be attached to the individual terms in the query representation. This is this is a generalization of document and query representation and processing.

Citation by others: Bordogna et al. [1993].

6.4 Lee et al. 1993

Citation:

LEE, J. H., KIM, W.Y., KIM, M. H. AND LEE, Y. J. 1993. On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework. *Proceeding: SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 291-297*

Problem: The authors state that the boolean retrieval systems have been the commonly used information retrieval system because of the efficient retrieval results and easy query structure. But the ranking is not supported and similarity coefficients cannot be calculated between queries and documents in traditional boolean retrieval systems. The fuzzy set model and the extended boolean model have been suggested providing the ranking function to the traditional boolean system and they are logical extensions of Boolean model because they reduce to boolean model when document term weights are restricted to zero or one. But still two problems exist: incorrect ranked output in some case and complex computation on boolean operators.

Previous work: The authors refer to previous work by Bookstein, [1980], Salton [1989].

Shortcomings of Previous work: The authors state that Bookstein's works on the fuzzy set model generate incorrectly ranked output in certain cases because the MIN and MAX operators have properties adverse to retrieval effectiveness. T-Operators in the fuzzy set theory have the Single Operand Dependency Problem and Negative Compensation Problem. Salton's work on the extended Boolean model has solve the problem of the former by apply the E_{AND} and E_{OR} operators; however, it suffers from the complex computation.

New Idea/Algorithm/Architecture: The authors give the T-operators and the corresponding operator graphs, and also, the averaging operators and corresponding operator graphs. Then using the graphs to describe that one pair of the Averaging operators of Fuzzy Sets Model and the operators of Extended Boolean Model overcomes the single operand dependency and negative compensation problems. These operators are defined as positively compensatory operators.

Experiments Conducted: The authors use two different document collections covering items the ISI collection and the CACM collection. They compare the precision of each operator such as T-operator and average operator to find the effectiveness of them. The authors use the document term weights to evaluate rank documents in Extended Boolean Retrieval Framework. "The weight of document is normalized as follow: W_{ik} = (TF_{ik}/maximum TF in document i)* (IDF_k maximum/IDF in document i) where Inverse Document Frequency define as IDF and Term Frequency define as TF."

Results: The authors state that positively compensatory operators provide higher retrieval effectiveness than the others. One pair of the Averaging operators of Fuzzy Sets Model achieves similar retrieval effectiveness and higher retrieval efficiency in comparison with the operators of Extended Boolean Model.

Claims: The author claim that the extended Boolean model has overcome the single operand dependency problem of the fuzzy set model by developing the operators for the evaluation of the AND and OR operations.

Citation by others: Pohl et al. [2012].

6.5 Lee 1995

Citation:

LEE, J. H. 1995. Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval. *Computer Science Technical Reports*. Cornell University Ithaca, NY, USA.

Problem: The author states that each document is indexed with a set of keywords or terms, and each query contains terms connected with the Boolean operators AND, OR and NOT in the traditional Boolean model. However, the traditional model does not provide a document ranking function because it cannot compute similarity coefficients between query and documents and this function reflects the relevance between query and documents that has the same objective with term weight.

Previous work: The author refers to previous work by Sachs [1976], Radecki [1979] and Buell [1980].

Shortcomings of Previous work: The author states that MIN and MAX operators

have been developed to support ranking function in the past for Boolean retrieval system. However, they do not correspond well with human behavior for the calculation of query-document similarities and lead the fuzzy set model to generate incorrect ranked output in some cases.

New Idea/Algorithm/Architecture: The author analyzes the behavioral aspects of different operators for AND and OR operations and the four important properties of retrieval effectiveness: single operand dependency, negative compensation, double operand dependency and unequal importance. The author describes them through examples and the four important could decrease retrieval effectiveness in some circumstance. The author also suggests that the properties of positive compensation retrieval and equal importance may help retrieval effectiveness. The author defines an operator class called n-ary soft Boolean operators that is suitable for achieving high retrieval effectiveness.

Experiments Conducted: The author evaluates the effectiveness of sixteen different operators that can evaluate AND and OR operations in extended boolean models. The author uses a new large data collection to evaluate the effectiveness of the different operators in extended boolean models. The large data collection includes one of the TREC sub-collections which is called WSJD2. The author also uses document term weights to calculate document values and two famous weighting schemes have been used which are Fox-weights and INQUERY-weights.

Results: The author states that the experiment's results suggest that the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem. The double operand dependency problem as well as the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem. INQUERY-weights give better retrieval effectiveness to the fuzzy set and Waller-Kraft models than Fox-weights. Network Boolean has the positively compensatory property in some operand values. The effectiveness of the p-norm model is slightly better than the vector space model for the well-formulated Boolean query.

Claims: The author claims that this paper does not present optimal operators, but the properties being analyzed can be used as a high base line to approach optimal operators.

Citation by others: Pohl et al. [2012].

6.6 Monz 2007

Citation:

MONZ, C. 2007. Model Tree Learning for Query Term Weighting in Question Answering. *Lecture Notes in Computer Science*, Vol. 4425, 589-596.

Problem: The author states that procedure to find an answer may be stuck when the retrieval system failed to find relevant information. Therefore, the effectiveness of retrieval in a question answering system is important and the formulation of the queries has dominant influence in those retrieval components, especially in Boolean retrieval. Using the wrong query terms may lead a processing retrieval into a wrong way.

Previous work: The authors refer to previous work by Brill et al. [2002].

Shortcomings of Previous work: The author states that Brill worked on retrieving the documents which match the query string or boolean query but do not solve the weight term problem.

New Idea/Algorithm/Architecture: One of the author's ideas is to use different query structures to discriminate terms which have positive or negative influence in the effectiveness of question retrieved. By ranking different well performing query variants and calculating the term weight (presence and absence weights) in the higher ranked query variants, the positive or negative gain can be found. The other approach used to solve the problem is derived from the first approach: the same term can be positive or negative gain in different queries but cannot give a term weight which did not occur. By using the representative features which have the term's information in the question instead the actual terms. In addition, use of a decision tree machine learning method to learn the query term's usefulness for query formulation and then give the weight to the query term.

Experiments Conducted: The author first used TREC-9, TREC-10, and TREC-11 data sets to compare different ways of choosing terms from the questions and try to identify the optimal query term. And then use different enhanced calculation method to analysis term weight.

Results: The author states that the query structure has a dominant influence in process retrieval system from the comparison experiments' results and better query structure can be performance better. From all the experiments, there are two factors affecting the query term weight: the term frequency in a documents and the retrieved frequency.

Claims: The author claims that his work demonstrates the importance to learn the query term weight in query structure as well as the term selection will have dominant influence in the performance of the retrieval components that confirmed by decision tree. The author is also confident about using the approach in the paper will benefit the question answering systems.

Citation by others: Lioma and Blanco [2009].

6.7 Pasi 1999

Citation:

PASI, G. 1999. A logical formulation of the Boolean model and of weighted Boolean models. *In Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems*, London, UK, 1-11.

Problem: The author states that in order to model IR in the logical framework there is a need for a more general formal discipline. It is necessary to analyse the role of logic in IR by defining a model of information retrieval based on modal logics, which provides a general framework to define pre-existing IR models. The query evaluation process of the Boolean model and of weighted Boolean models will be exploited by analyzing the role of logic as a formal basis. The analysis will give better understanding of some query weight semantics.

Previous work: The author does not refer to any of my selected paper as their related work.

Shortcomings of Previous work: No short comings of previous work were mentioned by the author.

New Idea/Algorithm/Architecture: The author gives a formulation of the boolean model that expresses the evaluation structure of the boolean query. Fuzzy implications can be employed to generalize the logical interpretation of the boolean model and it is necessary to describe the extension of the boolean model. An extended boolean model gives a weighted indexing function that evaluating a query term and the index term weight and it is interpreted as the degree of relevance of document with respect to query term. To enrich the expressiveness of the boolean query language, numeric query weights have been introduced as an extension of the basic selection criteria, which become then pairs term-weight.

Experiments Conducted: The author uses QT (t, q) as the terms appearing in a given query q and IT (t, d) as the index terms belonging to the representation of document d. So, the logical interpretation of a Boolean query is the formal expression of the constraint imposed by a query term: $QT(t, q) \rightarrow IT(t, d)$. "The expression representing the query evaluation structure of the Boolean query q = (t_1 AND t_2 OR (NOT t_3)) is the following: $(QT(t_1, q) \rightarrow IT(t_1, d) \wedge QT(t_2, q) \rightarrow IT(t_2, d)) \vee (\neg QT(t_3, q) \rightarrow IT(t_3, d))$." There is an extension of the logical interpretation of the Boolean model to weighted Boolean models. The IMP and QW is the importance of the index terms in document and in query representations. "A query of the type $q = < t_1, w_1 >$ AND $< t_2, w_2 >$ OR $< t_3, w_3 >$, has an evaluation which is formally expressed by the following logical expression: $((QW(t_1, q) \rightarrow IMP(t_1, d)) \wedge (QW(t_2, q) \rightarrow IMP(t_2, d))) \vee (QW(t_3, q) \rightarrow IMP(t_3, d))$."

Results: In the interpretation, the weighted query is a part of the formulation. The constant symbols are terms in the formulation and the logical connectives correspond to the Boolean connectives. The choice of the implication operator is important in the formalization, as it is strictly connected with the semantics of the query term-weight.

Claims: The author claims that the approach is based on the following considerations: terms are the most essential elements which evaluate the relevance in a query; the natural logical interpretation of the boolean model is important and the degree of relevance related to the truth of the given interpretations. The author also claims that the most important part of this approach is to make the bottom-up structure of the query evaluation procedure clear and the implication connective is employed to express limitation controlled by a query term on the document representations.

Citation by others: Latiri et al. [2003].

6.8 Patro and Malhotra 2005

Citation:

PATRO, S. AND MALHOTRA, V. 2005. Characteristics of the Boolean Web search query: Estimating success from characteristics. *In Proceedings of First*

International conference on WEB Information Systems and Technologies (WEBIST 2005), May 26-28, 2005, Miami, Florida.

Problem: The authors state that whether the query will successfully satisfy the users' requirements are unknown. Similarly, whether the Boolean query is efficient or not in finding the location of the best documents is unknown.

Previous work: The authors do not refer to any of my selected paper as their related work.

Shortcomings of Previous work: No short comings of previous work were mentioned by the authors.

New Idea/Algorithm/Architecture: The authors state that this paper relates characteristics of the search results which meet the user's requirement. A program which compares the performance of humans and the search queries synthesized performed better than humans and gives an objective judgment of the search queries.

Experiments Conducted: The author states that use the precision and recall to measure the quality of queries is the common method. When search query, use the number of relevant documents first returned 20 links as the precision measurement. The authors use the same range as precision to identify recall. The authors collect data by using volunteers. Volunteers are students have different topic area and some sample relevant documents. After they use the same search engine to search their topic, they revised their query. Then compare the volunteer's query and the synthesized query. The authors use figures to show the relationship between recall and precision of the volunteer queries; the relationship between recall and precision after the queries' quality changed; the precision as function of terms in query and number of attempts to improve query and relationship between number of terms and precision.

Results: A good query returns higher values for precision and good recall which can be used as evaluation results of the method applied to the boolean model. A good predictor of a successful Boolean web search query is four terms and above average recall.

Claims: The authors state that they found the characteristics of good queries which may give good sample of the query successfully achieved by users and their requirements.

Citation by others: Patro, Malhotra and Johnson [2007].

6.9 Pohl et al. 2012

Citation:

POHL, S., MOFFAT, A. AND ZOBEL, J. 2012. Efficient Extended Boolean Retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, 24(6), 1014-1024.

Problem: Boolean queries have the disadvantage of being harder to formulate than ranked queries. They have the drawback of generating answer lists of unpredictable length and changes in the query that appear to be small might result in disproportionately large changes in the size of the result set. The queries of Extended Boolean Retrieval (EBR) models, such as the p-norm model, queries are slow to evaluate, because of their complex scoring functions and none of the computational optimizations available for ranked keyword retrieval have been applied to EBR.

Previous work: The authors do not refer to any of my selected paper as their related work.

Shortcomings of Previous work: No short comings of previous work were mentioned by the author.

New Idea/Algorithm/Architecture: The authors describe a scoring method for EBR models and adopts ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions. The authors also present the p-norm EBR model as an instance of such models and that performance gains can be attained that are similar to the ones available when evaluating ranked queries. Term-independent bounds are proposed in this paper, which complement the bounds obtained from max-score. It can be employed in the wand algorithm, also reducing the number of score evaluations.

Experiments Conducted: The authors evaluated the efficiency of their methods on a large collection of biomedical literature using queries and results derived from real searches. Three query sets were used against this collection: 50 simple, short PUBMED queries consisting of a Boolean conjunction only, 50 structured queries containing both conjunctive and disjunctive operators at least once in each query sampled from the same query log, 15 complex queries. Properties of these queries are summarized. The authors counted the number of scored documents with scores below the entry threshold and above the entry threshold.

Results: The authors state that the proposed scoring method is often being faster and it significantly reduces the number of candidate documents scored, postings processed, and execution times, for all query sets. Term-independent bounds method for short-circuiting candidate document scoring reduce the number of score calculations, especially on simpler queries and when combined with max-score. The query execution times of boolean execution will be faster, however, it must be remembered that the result of a Boolean query is of indeterminate size.

Claims: The authors claim that optimization techniques developed for ranked keyword retrieval can be modified for EBR and this leads to considerable speedups. Term-independent bounds provide added benefit when complex scoring functions are used and it will as a mean for short-circuit score calculations.

Citation by others: Vignesh and Sivakumar [2013].

6.10 Wong et al. 1986

Citation:

WONG, S. K. M., ZIARKO, W., RAGHAVAN, V. V., AND WONG, P. C.N. 1986. On extending the vector space model for Boolean query processing. *Proceeding: SIGIR '86 Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*. 175-185.

Problem: The authors state that the boolean retrieval systems do not apply the incorporating term correlations into the retrieval process. In another words, the weighted queries and documents is a problem for Boolean retrieval systems.

Previous work: The authors refer to previous work by Buell [1981].

Shortcomings of Previous work: The authors state that Buell's work on the strict Boolean retrieval systems has the problem that there is no provision for weights of importance to the terms both in the queries and documents. The representation is binary that lacks various index terms and the output is not ranked. In most cases, the AND connectives tend to be too restrictive.

New Idea/Algorithm/Architecture: The authors introduce a new information retrieval model, named Generalized Vector Space Model (GVSM). The new model solves the weights problem in the boolean model and the queries used in this model are seemed as an extended boolean expressions. In GVSM, a query is simply defined as a weighted vector sum of term vectors. However, this form of query does not help the user to explain clearly structure as can be done in boolean systems. Therefore, the most common language used to express query structure involves boolean logic and the query is considered to be a list of index term and weight pairs. This is the first important variation queries that are called the basic GVSM. The other one is the called unified GSVM which involves a scheme for expressing weighted Boolean queries as vectors. Queries are specified as a weighted boolean expression in which are connected by AND, OR and NOT operators.

Experiments Conducted: The authors compare the unified model with the p-norm model which was applied for extended boolean retrieval model. The authors use MEDLARS and CISI collections for experimental evaluation, because other collections used for information retrieval do not provide boolean queries. The standard recall and precision measures are used for comparing the performance of different models.

Results: After comparing these models, the authors find that both models handle weighted boolean queries, both reduce to VSM and strict Boolean retrieval models under certain conditions, p-norm model involves the parameter p, which has to be experimentally determined and the extended query language satisfies more algebraic properties under unified GVSM. The unified GVSM is closer to the strict Boolean model than it is to VSM which is means the roles of Boolean operators is rather strictly retained, however, p-norm model achieves more softening of the operators.

Claims: The authors claim that it would be advantageous to provide a prescription to handle boolean queries in the GVSM environment. The important part of basic GVSM is generalizing VSM to incorporate term correlations and document representation reduces to the vector sum of terms when terms are assumed to be orthogonal. The unified GVSM reduces to the strict Boolean retrieval model when each document is represented by its dominant atomic vector.

Citation by others: Wu et al. [2008].

7. REFERENCES

REFERENCES

BOOKSTEIN, A. 1980. Fuzzy Requests: An Approach to Weighted Boolean Searches. *American Society for Information Science.* 31(4), 240-247. (Exactly on topic)

BORDOGNA, G., CARRARA, P. AND PASI, G. 1991. Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27(1), 15-26. (Exactly on topic)

BORDOGNA, G. AND PASI, G. 1993. A fuzzy linguistic approach generalizing Boolean Information Retrieval: A model and its evaluation. *Journal of the American Society for Information Science*, 44(2), 70–82. (This paper refer to some of my selected papers and part of this paper is exactly on topic)

BUELL, D.A. 1982. An Analysis of some Fuzzy Subset Applications to Information Retrieval Systems. *Fuzzy Sets and Systems*, 7(1), 35-42. (This paper refer to some of my selected papers and part of this paper is exactly on topic)

KANTOR, P. B. 1981. The logic of weighted queries. *IEEE Transaction on Systems Men and Cybernetics*, 11(12), 816-821. (This paper refer to some of my selected papers and part of this paper is exactly on topic)

KRAFT, D. H. AND BUELL, D. A. 1983. Fuzzy sets and generalized Boolean retrieval systems. *International Journal of Man-Machine Studies*, 19(1), 45-56. (Exactly on topic)

KRAFT, D. H., BORDOGNA, G., AND PASI, G. 1994. An extended fuzzy linguistic approach to generalize boolean information retrieval. *Information Sciences – Applications*, 2(3), 119-134. (This paper refer to some of my selected papers and part of this paper is exactly on topic)

LATIRI, C. C., YAHIA, S. B. AND CHEVALLET, J. P. 2003. Query expansion using fuzzy association rules between terms. *In Knowledge discovery and discrete mathematics. International conference*, 231-242. (This paper refer to some of my selected papers and part of this paper is exactly on topic)

LEE, J. H., KIM, W.Y., KIM, M. H. and LEE, Y. J. 1993. On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework. *Proceeding: SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 291-297. (Exactly on topic)

LEE, J. H. 1995. Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval. *Computer Science Technical Reports.* Cornell University Ithaca, NY, USA. (Exactly on topic)

LIOMA, C. AND BLANCO, R. 2009. Part of Speech Based Term Weighting for Information Retrieval. *Advances in Information Retrieval*, 5478, 412-423. (This paper refer to some of my selected papers)

MONZ, C. 2007. Model Tree Learning for Query Term Weighting in Question Answering. *Lecture Notes in Computer Science*, Vol. 4425, 589-596. (Exactly on topic)

PASI, G. 1999. A logical formulation of the Boolean model and of weighted Boolean models. *In Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems*, London, UK, 1-11. (Exactly on topic)

PATRO, S. AND MALHOTRA, V. 2005. Characteristics of the Boolean Web search query: Estimating success from characteristics. *In Proceedings of First International conference on WEB Information Systems and Technologies (WEBIST 2005)*, May 26-28, 2005, Miami, Florida. (Exactly on topic)

PATRO, S., MALHOTRA, V., JOHNSON, D. 2007. An Algorithm to Use Feedback on Viewed

Documents to Improve Web Query. *Web Information Systems and Technologies*, 177-189. (This paper refer to some of my selected papers)

POHL, S., MOFFAT, A. and ZOBEL, J. 2012. Efficient Extended Boolean Retrieval. Knowledge and Data Engineering, *IEEE Transactions on*, 24(6), 1014-1024. (Exactly on topic)

SALTON, G. AND WU, H. 1980. A term weighting model based on utility theory. *Proceeding: SIGIR '80 Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, 9-22. (This paper refer to some of my selected papers)

VIGNESH, U. AND SIVAKUMAR, M. 2013. Implementing High Performance Retrieval Process by Max-Score Ranking. *Journal of Computer Engineering (IOSR-JCE)*, 8(5), 28-33. (This paper refer to some of my selected papers)

WONG, S. K. M., ZIARKO, W., RAGHAVAN, V. V., Wong, P. C.N. 1986. On extending the vector space model for Boolean query processing. *Proceeding: SIGIR '86 Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*. 175-185. (Exactly on topic)

WU, H. C., LUK, R. W. P., WONG, K. F., KWOK, K. L. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3). (This paper refer to some of my selected papers and part of this paper is exactly on topic)