J Naveen Kumar *et al*, International Journal of Computer Science and Mobile Applications,

Vol.1 Issue. 5, November- 2013, pg. 38-43 **ISSN: 2321-8363**

EFFECTIVE EFFICIENT BOOLEAN RETRIEVAL

J Naveen Kumar[1], Dr. M. Janga Reddy[2]

[1] jnaveenkumar6@gmail.com, [2] pricipalcmrit@gmail.com

[1] M.Tech Student, Department of Computer Science, CMR Institute of Technology, Medchal, Hyderabad

[2] Professor and Principal, Department of Computer Science, CMR Institute of Technology, Medchal, Hyderabad

ABSTRACT

The conventional Boolean retrieval system does not provide ranked retrieval output because it cannot compute similarity coefficients between queries and documents. Extended Boolean retrieval (EBR) models were proposed nearly three decades ago, but have had little practical impact, despite their significant advantages compared to either ranked keyword or pure Boolean retrieval. The Boolean retrieval model contrasts with ranked retrieval models such as the vector space model in which users largely use free text queries , that is, just typing one or more words rather than using a precise language with operators for building up query expressions, and the system decides which documents best satisfy the query. Despite decades of academic research on the advantages of ranked retrieval, systems implementing the Boolean retrieval model were the main or only search option provided by large commercial information providers for three decades until the early 1990s (approximately the date of arrival of the World Wide Web). However, these systems did not have just the basic Boolean operations (AND, OR, and NOT) which we have presented so far. A strict Boolean expression over terms with an unordered results set is too limited for many of the information needs that people have, and these systems implemented extended Boolean retrieval models by incorporating additional operators such as term proximity operators. A proximity operator is a way of specifying that two terms in a query must occur close to each other in a document, where closeness may be measured by limiting the allowed number of intervening words or by reference to a structural unit such as a sentence or paragraph.

Keywords- Document-at-a-time; efficiency; extended Boolean retrieval; p-norm; query processing

I. INTRODUCTION

Information Retrieval (IR) systems represent, store and retrieve the input information, which is likely to include the natural language text of documents or of document and abstracts. The output of IR systems in response to a search request consists of references. These references are intended to provide the system users with the information about items of potential interest.

In general the retrieval of the resultant relevant document list from the large dump of document s is a greatest task for the IRS systems to satisfy the general set of users using their search pattern criteria and based on the indexing.

But, In case of the Effective Extended Boolean Retrieval Model, is based on the exact relevant Pattern Match technique which search for the existence of the exact pattern throughout the document and give the resultant relevant set of documents that has the Max Score.

This Idea of Implementation is based on the construction of the efficient query tree with the identifiers that represent the search pattern from the top to the Bottom. The search try to identify these pattern from the root follow to the bottom in search of the relevant search pattern and when it identifies the complete pattern at the particular level then it process all the nodes down in the next level and go for calculation of the Max Score.

This paper implements the Max Score which proposes the idea of frequency the pattern that exists in the document based on the relevance content in every document in the document set.

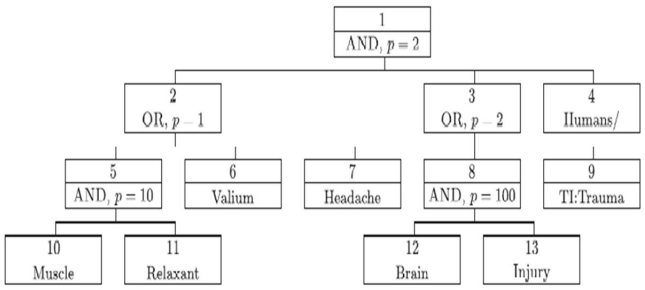


Fig. 1. Example query tree with assigned node identifiers, p-values, terms, and their document frequencies.

A major role of IR systems, however, is not just to generate a set of relevant references, but to help determine which documents are most likely to be relevant to the given requirements. IR systems should present to users a sequence of documents ranked in decreasing order of query-document similarities. Users are able to minimize their time spent to find useful information by reading the top-ranked documents first. Boolean retrieval systems have been most widely used among commercially available IR systems due to efficient retrieval and easy query formulation. In conventional Boolean retrieval systems, however, document ranking is not supported and similarity coefficients cannot be computed between queries and documents. The fuzzy set model and the extended Boolean model have been proposed to overcome this problem. They are logical extensions of the Boolean retrieval system because they reduce to the Boolean model when document term weights are restricted to zero or one.

Id.	Term	Frequency
4	Humans/	9,912,283
6	Valium	603
7	Headache	41,671
9	TI:Trauma	36,313
10	Muscle	461,254
11	Relaxant	7,244
12	Brain	733,025
13	Injury	274,994

Advantages of Boolean retrieval include:

Complex information need descriptions: Boolean queries can be used to express complex concepts;

Composability and Reuse: Boolean filters and concepts can be recombined into larger query tree structures;

Reproducibility: Scoring of a document only depends on the document itself, not statistics of the whole collection, and can be reproduced with knowledge of the query;

Scrutability: Properties of retrieved documents can be understood simply by inspection of the query; and

Strictness: Strict inclusion and exclusion criteria are inherently supported, for instance, based on metadata.

II. IMPLEMENTATION ISSUES

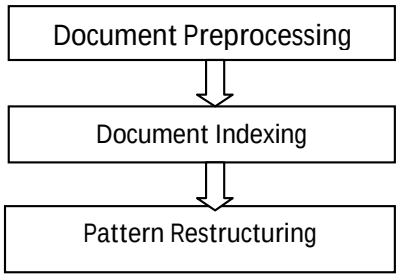
This half implementation is with respect to the Search Engine Maintenance and development part of the paper

Document Pre-processing: In the first module the documents contain sentences. The sentences are in the unstructured manner. The module converts sentences to structured sentences with index. This process is applied on the existing corpus.

Document Indexing: In this module each sentence of a document is made up with different words.
Example: S1= {w1, w2, w3…….wn}
The module splits all the indexed sentences by words.

Pattern Restructuring: In this module, the words will be presented in the document in different forms such as present, past, future etc…The words has to be n-gramered to find out the possible equivalence of root words. The root words can be grouped together (or) clustered for special group of interests.

Example: {—cricket□, —football□} can be grouped together to special interests called —sports□ category. Identifying group of words of similar category can have relationship. Building the relational words together is called word-net.

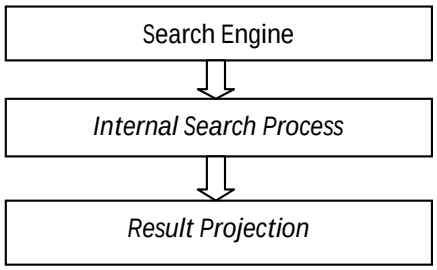


The second half Part is based on the users part of the paper where the search engine behaves based on the users search Pattern for the result document list.

Search Engine: The word-net is a semantic relational network. The word-net is store in the database as Parse Tree Data Base(PTDB). The module provides an interface to the user to search the PTDB of the corpus. The user's query will be in the form of natural language (or) can be with stop words.

Internal Search Process: In this module, user's query has to be pre-processed against stop words elimination. The query words have to be n-grammed for possible root words.

Result Projection: In this module, all the n-grammed words may not be the root words. Find out the possible root words for each query word. Find the semantically words for each word of query root word. Find the appropriate Tag with their relevancies (or) Frequencies.



III. CONCLUSION

Having noted that ranked keyword querying is not applicable in complex legal and medical domains because of their need for structured queries including negation, and for repeatable and scrutable outputs, we have presented novel techniques for efficient query evaluation of the p-norm (and similar) extended Boolean retrieval model, and applied them to document-at-a-time evaluation. We showed that optimization techniques developed for ranked keyword retrieval can be modified for EBR, and that they lead to considerable speedups. Further, we proposed term-independent bounds as a means to further short-circuit score calculations, and demonstrated that they provide added benefit when complex scoring functions are used. A number of future directions require investigation. Although presented in the context of document-at-a-time evaluation, it may also be possible to apply variants of our methods to term-at-a-time evaluation. Second, to reduce the number of disk seeks for queries with many terms, it seems desirable to store additional inverted lists for term prefixes (see, for example, Bast and Weber), instead of expanding queries to hundreds of terms; and this is also an area worth exploration. We also need to determine whether or not term-dependent bounds can be chosen to consistently give rise to further gains. As another possibility, the proposed methods could further be combined and applied only to critical or complex parts of the query tree. Finally, there might be other ways to handle negations worthy of consideration. We also plan to evaluate the same implementation approaches in the context of the inference network and wand evaluation models. For example, it may be that for the data we are working with relatively simple choices of term weights—in particular, strictly document-based ones that retain the scrutability property that is so important—can also offer good retrieval effectiveness in these important medical and legal applications.

REFERNCES

[1] S. Karimi, J. Zobel, S. Pohl, and F. Scholer, ―The Challenge of High Recall in Biomedical Systematic Search,‖ Proc. Third Int'l Workshop Data and Text Mining in Bioinformatics, pp. 89-92, Nov. 2009.

[2] J.H. Lee, ―Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval,‖ Technical Report TR95-1501, Cornell Univ., 1995.

[3] G. Salton, E.A. Fox, and H. Wu, ―Extended Boolean Information Retrieval,‖ Comm. ACM, vol. 26, no. 11, pp. 1022-1036, Nov. 1983.

[4] J.H. Lee, W.Y. Kin, M.H. Kim, and Y.J. Lee, ―On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework,‖ Proc. 16th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 291-297, 1993. POHL ET AL.:EFFICIENT EXTENDED BOOLEAN RETRIEVAL 1023

Vol.1 Issue. 5, November- 2013, pg. 38-43 **ISSN: 2321-8363**

Authors' Profile

First A. Author: ***J Naveen Kumar*** received B.Tech Degree in Computer Science and Engineering from JNTUH in the year of 2011. He is currently M.Tech student in the Computer Science Engineering from Jawaharlal Nehru Technological University (JNTUH), Hyderabad. And she is interested in the field of Data Mining.

Second Author: ***Dr. M. Janga Reddy*** working as Principal, CMRIT, Hyderabad. Since inception i.e. from 2005, he has been working as Professor of CSE and Principal. Ratified by JNTUH, Hyderabad (2010). Teaching Experience: 20 years. International/National Research Papers: 56. Life-Time member of ISTE, IETE, CSI. Nominated as member, UGC – Academic Staff College, JNTUH Academic Council by Vice-Chancellor. Honoured as best teacher by Lions Club, Gymkhana, RR Dist., for the A.Y. 2011-12. Guided many B.Tech and M.Tech students in their project work. Supervises Ph.D. scholar. Editorial member for several reputed Journals.

