ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.7

Boolean Retrieval Model for Indexed Documents

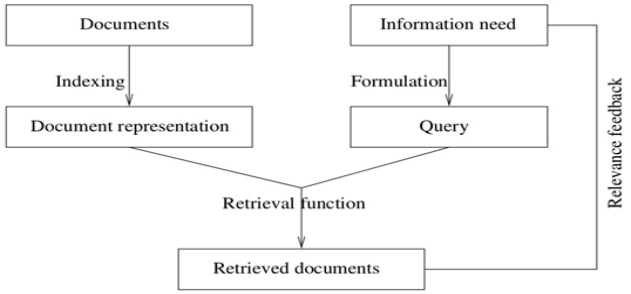
Deepika Sharma[1], Jay Kumar[2]
M.Tech Student[1], Assistant Professor[2]
Department of CSE
RGEC, Meerut, India

Abstract:
Information of all the different domains are available online in the form of hyper text in web pages. Peoples according to their need consulting different web sites to fetch information. It is very difficult to remember the names of the websites for a specific domain for which the user wants to search. So a search is a system which mines information from the WWW and present it to the user according to its query. Information retrieval system (IRs) works for search engine arranges the web documents systematically and retrieves the result according to the user query. In this paper a model is proposed based on boolean retrieval which retrieves the results according to the according to the Boolean operation specified within the terms of the search query. Also the proposed model is capable to store large indexes.

Keywords: Search engine, Information Retrieval, indexes, Boolean retrieval.

1. INTRODUCTION

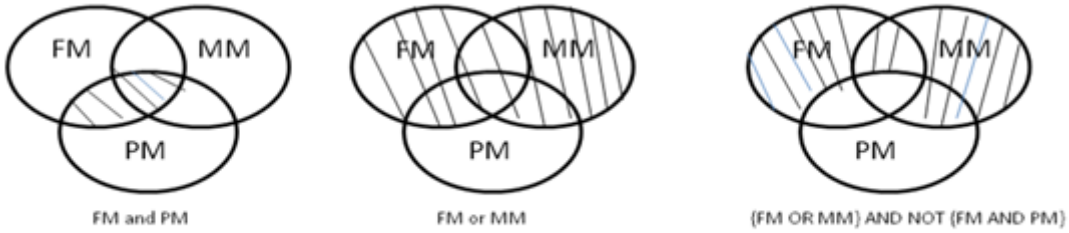
The meaning of the term *information retrieval* can be very broad. As an academician theory, Information retrieval (IR) is retrieving unstructured (text) documents that satisfy an information need from within large collections of documents stored on computers. According to the above definition information retrieval can be used by librarians, paralegals, and similar professional searchers. Millions of peoples are using information retrieval in web search engines for finding their desired information. Information retrieval is fast becoming the dominant form of information access. IR also deals with other kinds of data and information problem other than mentioned above. Unstructured data means information that is semantically not correct and cannot be used by the computer structure. It is the opposite of structured data, the example of structured data is RDBMS, which maintain product inventories and personnel records of small companies. In reality, almost no data are truly "unstructured". This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has specific format, such as headings, paragraphs and footnotes, which are highlighted in the document (such as the coding underlying web pages). IR is also used to facilitate "semi structured" search such as finding a document where the title contains c++ and the body contains templates. The area of information retrieval also covers supporting users in browsing or filter document repositories or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of titles, standing information necessity, or other categories (such as accuracy of texts for age group of different peoples), classification is deciding of which classes, if any, each of a set of documents belongs to. It is often accomplished by manually classifying documents and then uses that idea to classify new documents automatically. Information retrieval systems differentiate by the scale at which they operate, and is useful for distinguishing three prominent scales. In *web search*, provides facilities to mine millions of web documents stored on various computers located geographically at distinguish locations. Distinctive issues are required for document indexing, being able to build systems that work efficiently at enormous scale, and handling particular aspects of the web, such as taking the benefit of hypertext of site to boost their search engine rankings, given the commercial importance of the web. At the other extreme is *personal information retrieval*. In the last few years, operating systems embed information retrieval. Email programs provides search as well as text classification: they at least provide a spam (junk mail) filter, and commonly also provide either manual or automatic means for classifying mail so that it can be placed directly into particular folders. There are various distinguish issues such as handling the broad range of document types on a personal computer, and making the search system maintenance free and sufficiently lightweight in terms of startup, processing, and disk space usage that it can run on one machine without annoying its owner. In between is the space of *business, institutional, and domain-specific search*, where retrieval might be provided for collections of corporation's internal documents, a database of patents, or research articles on agri informatics. In this case, the documents are stored on centralized file systems and one or a more machines will provide for search over the collection. The figure below shows retrieval-



IRs perform the following activities to achive its goal-
1.In indexing the documents are arranged with respect to the terms in the document.
2.Removal of unnecessary words, frequently used words which have less contribution in giving the watage to the document with respect to terms of the document.
3.Fetching of documents according to the user query.

2. BOOLEAN RETRIEVAL MODEL

The *Boolean retrieval model* is a model for information in which any user query can be posted which is in the form of a Boolean expression of terms, that is, the terms are attached with the operators AND, OR, and NOT. The model views each document as just a set of words. The Basic assumptions of IR are collection of fixed collection of documents and motive is to retrieve documents with information that is relevant to the user's information need and helps the user complete a task. The retrieval model considers each document as relevant or irrelevant according to the user query. The figure below shows the visualization of Boolean retrieval model among the three set of documents.



FM: File Management, MM: Memory Management, PM: Process Management

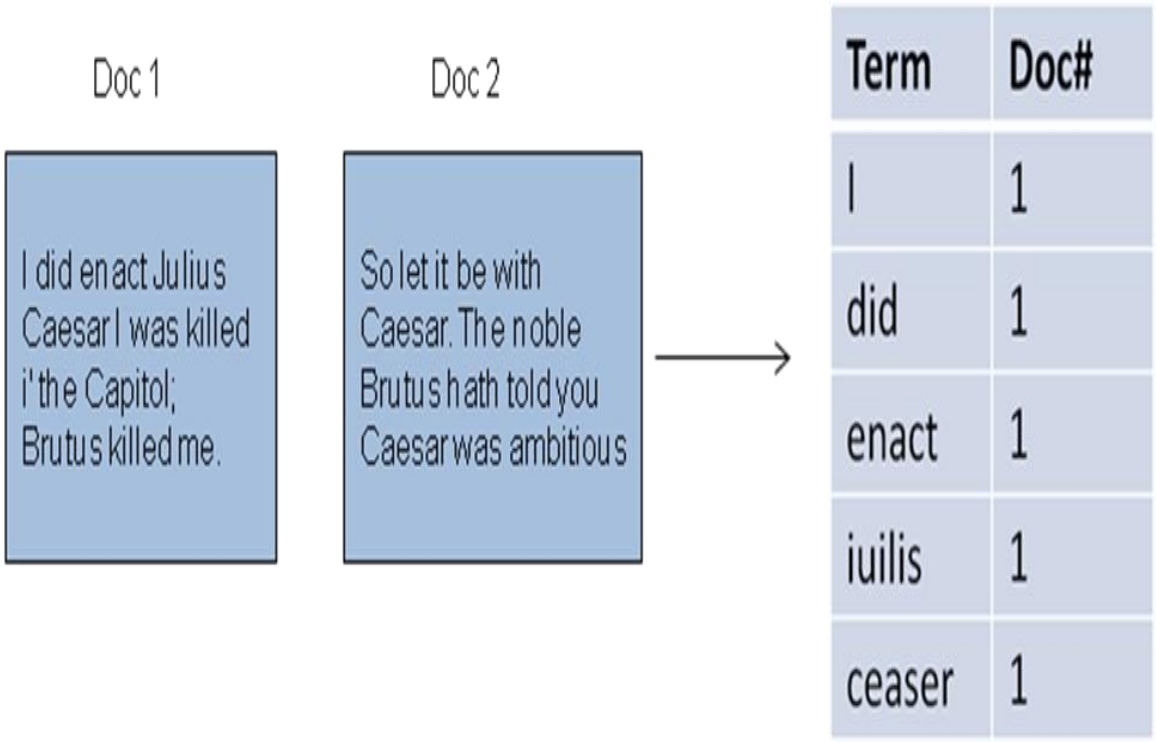
2.1 INDEX CREATION IN BOOLEAN RETERIVAL MODEL

Let us now consider a more realistic scenario, simultaneously using the opportunity to introduce some terminology and notation. Suppose we have N = 1 million documents. By *documents* we mean whatever units we have decided to build a retrieval system over. They may be chapters of a book. The group of documents over which retrieval operation is performed is known as the (document) *collection*. It is also referred to as a *corpus* (a *body* of texts). Suppose each document contains about 2000 words long (4–5 book pages). If we assume an average of 5 bytes per word including spaces and punctuation, then the document collection is of about 8 GB in size. Typically, there might be about M = 600,000 distinct terms in these documents. There is nothing special about the numbers we have chosen, and they might vary by an order of magnitude or more, but they give us some idea of the dimensions of the kinds of problems we need to handle. This idea is central to the first major concept in information retrieval, the inverted *index*. The name is actually redundant: an index always maps back from terms to the parts of a document where they occur. Nevertheless, *inverted index*, or sometimes *inverted file*, has become the standard term in information retrieval.

2.2 Steps of indexer

2.2.1. Token sequence

It generates Sequence of (Modified token, Document ID) pairs.



2.2.2 Sort by terms

Sort the sequence alphabetic vise.

Term	Doc#
I	1
did	1
enact	1
iuilis	1
ceaser	1

→

Term	Doc#
I	1
did	1
enact	1
ceaser	1
iuilis	1

2.2.3. Dictionary and postings

Multiple term occurrences in a single document are merged. Split into Dictionary and Postings and Document frequency information is added. The process is shown in the figure 2.2.3

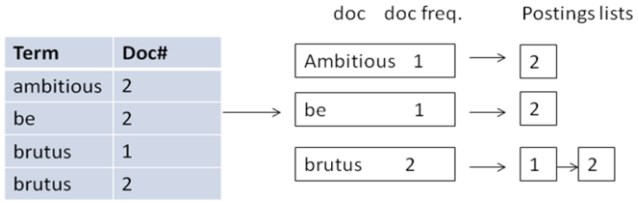


Figure. 2.2.3 Dictionary and postings

2.3 Query processing

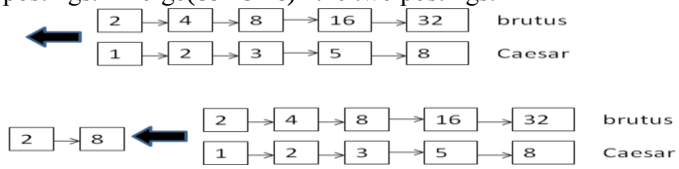
After the posting list created for the terms, then the query is processed to find the resultant documents from the postings. For example-

Consider processing the query:

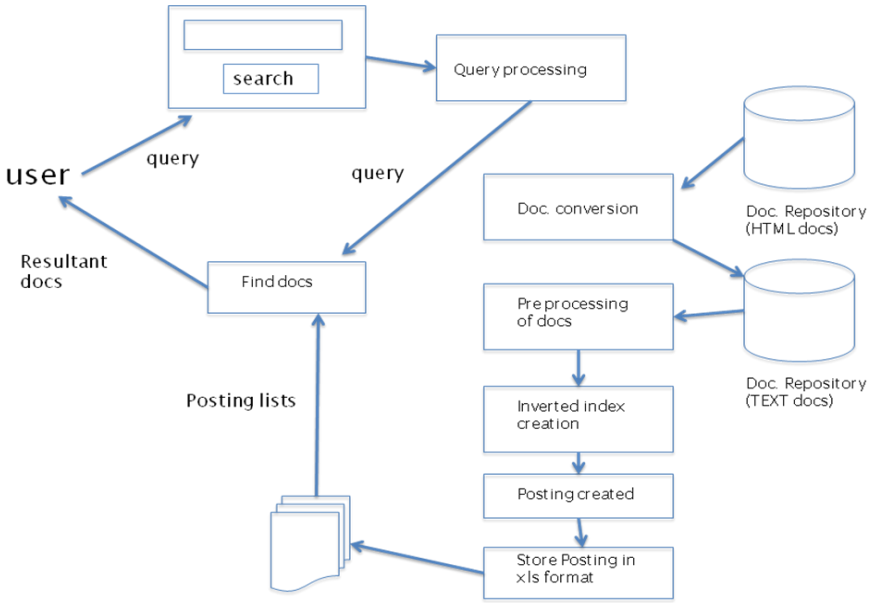
Brutus AND Caesar

Locate ***Brutus*** in the Dictionary;

Fetch its postings. Locate *Caesar* in the Dictionary; Retrieve its postings. "Merge(combine)" the two postings:

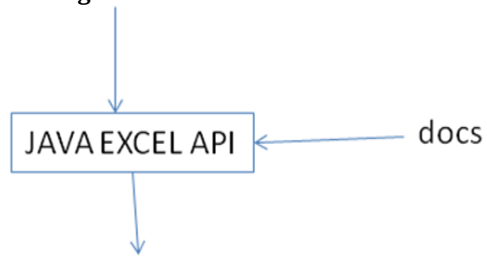


3. ARCHITECTURE OF PROPOSED MODEL



After the user supplies the query, the query is processed and unnecessary terms will be removed from the query, the resultant query only contains the keywords with the proper Boolean operator. The documents available in the repository in html form is converted into text documents and preprocessed for removing cure words and meaningless words from the document, after conversion of the document into text document its size decreases by removing unnecessary tags from the html document. After that one by one document from the text repository is fetched and the inverted indices of the terms of the document is created now the postings created are stored in the excel file because storing the postings in excel is efficient as the size of the posting lists increases if the number of documents are more and storing these postings list in any other data structure is not efficient. Finally the created postings are merged and the resultant documents are generated according to the query.

3.1 Storage of indexes



term	Doc id's			
process	1	2	4	10
SJF	2	5	7	9

The indexes created are stored in the excel sheet by using the java excel API, which provides a large storage to the indexes.

5. CONCLUSION

At last we make a conclusion that, information retrieval is a process of finding and fetching the knowledge based information from cluster or collection of documents. Boolean retrieval model used for information fetch is more accurate as compared to other retrieval models. The model creates the inverted indexes of terms and docs, on which boolean operation can be applied easily and show accurate result.

6. REFERENCES

[1] M.François Sy, S.Ranwez, J.Montmain,"User centered and ontology based information Retrieval system for life sciences", BMC Bioinformatics,2105.

[2] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", IJCER, sep 2012, Vol. 2 Issue. 5, , PP: 1443-1444,.

[3] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines echniques",2010,Al- Satil journal,PP: 55-92.

[4] D.Hiemstra,P. de Vries, "Relating the newlanguage models of information retrieval to the traditional retrieval models", published as CTIT technical report TR-CTIT-00-09, May 2000.

[5] Djoerd Hiemstra, "Information Retrieval Models", published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, November 2009,Ltd., ISBN-13: 978-0470027622.

[6] Christos Faloutsos, Douglas W. Oard, "A Survey of Information Retrieval and Filtering Methods", CS-TR-3514, Aug 1995. "Algorithms for Information Retrieval – Introduction", Lab module 1.

[7] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval",2009, ACM Press, ISBN: 0-201-39829-X.

[8] S.E. Robertson and K. Sparck Jones. "Relevance weighting of search terms. Journal of the American Society for Information Science", 1976, 27:129–146.

[9] G. Salton and M.J. McGill, "editors. Introduction to Modern Information Retrieval". McGraw-Hill ,1983.

[10] H. Turtle, "Inference Networks for Document Retrieval". Ph.D. thesis, Department of Computer Science,University of Massachusetts, Amherst, MA 01003. Available as COINS Technical Report 90-92, 1990.

[11] C. J. van Rijsbergen. "Information Retrieval. Butterworths", London,1979. [12] T. Strzalkowski, L. Guthrie, J. Karlgren, J. and et. "Natural language information retrieval: TREC-5 report". In Proceedings of the Fifth Text REtrieval Conference (TREC-5), 1997.