ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

Review of Indexing Techniques in Information Retrieval

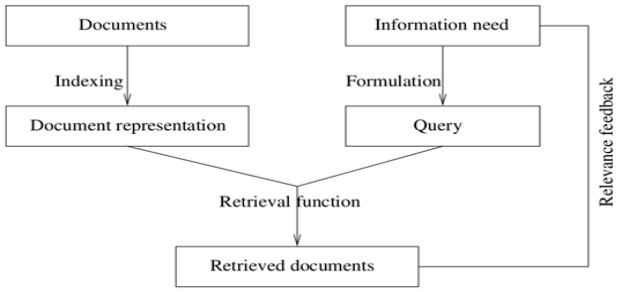
Ekta Chauhan[1], Dr. Amit Asthana[2]
M.Tech Student[1], HOD[2]
Department of CS
Shubharti University, Meerut, India

Abstract:
The way to avoid linearly scanning the texts for each query is to index the documents in advance. Document indexing is a powerful technique to aid subsequent retrieval of documents from repositories that contain thousands of documents. Document might be indexed by their full-text content (so that they can be retrieved by any word in the content) or by metadata attached to the document such as a unique identifier, date of creation, or the main topic of the document. The basic steps in constructing a non positional index are: first make a pass through the collection assembling all term–docID pairs. Then sort the pairs with the term as the dominant key and docID as the secondary key. Finally, organize the docIDs for each term into a postings list and compute statistics like term and document frequency. For small collections, all this can be done in memory. This paper discusses various indexing techniques used in information-retrieval model and their evaluation criteria.

Keywords: information retrieval, doc ID, indexing

1. INTRODUCTION

Information Retrieval is a process Retrieving and Presenting various content object to the user relevant to his/her query from a standardized collection of objects from different sources or repositories. Web is the best resource of Information Retrieval Processes, where different techniques are used to give exact information needed by the users. Naive users are not much familiar with structural queries. Users submit short queries that do not consider the variety of terms used to describe a topic, resulting in poor recall power [5]. Searching on non standardized store of bulk document is highly difficult, where in indexing reduces the complexity of search process. Information Retrieval is process of Indexing is a process of identifying keywords to represent a document based on their contents. Indexing is very important phase of Information Retrieval System to create a search-able unit for the given query. Basically, indexing is performed by assigning each document with keywords or descriptive terms representing the document[1]. The assigned terms must reflect the content of the document to allow effective keyword searching. In automatic indexing, couple of trained people who are well with concept of the document participates in indexing process. Manual indexing is a time taking process and it requires huge manual hours to index a repository which grows day by day. Automatic text indexing which is much faster and less error-prone has become a common practice on big corpus. Research on English texts has shown that the retrieval effectiveness of automatic indexing is comparable to that of manual indexing [2][3]. A natural language query specifies the user's information need in a natural language sentence or sentences. A phrase query contains phrases representing concepts of interest to the user [4], then it requires to mind the language features before selecting indexing terms. Where in the language processing tools helps to identify the better indexing terms to represent whole object. In this case study the effect of various indexing techniques are observed on fixed length Telugu corpus. Majority of related work has been examined on various language corpus through literature survey in next chapter. In this paper we concentrated on how the indexing improves the retrieval performance. Various methods are adopted to index the items in this research. Indexing of items using semantic concept gives better representations. 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.Retrieving of documents according to the user query.

2. INDEX CREATION TECHNIQUES

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*. An index always maps from terms to the parts of a document where they occur. I*nverted index*, or sometimes *inverted file*, used as a standard term in information retrieval. The basic use of an inverted index is illustrated in Figure 2.1a. We keep a *dictionary* of terms (*vocabulary).* Then for each term, a list of records of documents prepared that contains the term. Each item in the list – which records that a term appeared in a document is called a *posting*. The list is then called a *postings list* (or inverted list), and all the postings lists taken together are referred to as *postings*. The dictionary has been arranged alphabetically and each postings list is sorted by document ID.

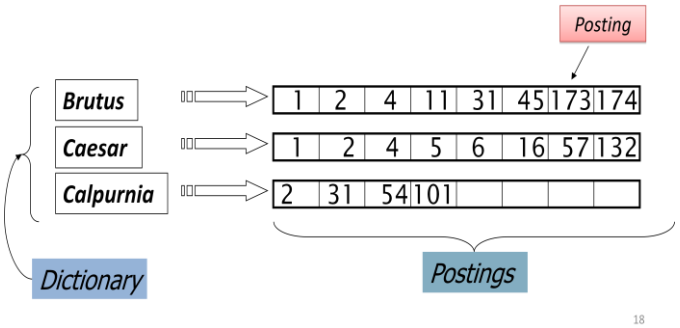


Figure.1. a Indexing of documents.

2.2 Latent semantic indexing

Latent semantic indexing (**LSI**) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSI begins by constructing a term-document matrix, A, to identify the occurrences of the m unique terms within a collection of n documents. In a term-document matrix, each term is represented by a row, and each document is represented by a column, with each matrix cell, a_{ij}, initially representing the number of times the associated term appears in the indicated document, tf_{ij}. This matrix is usually very large and very sparse. Once a term-document matrix is constructed, local and global weighting functions can be applied to it to condition the data. The weighting functions transform each cell, a_{ij} of , A to be the product of a local term weight, l_{ij}, which describes the relative frequency of a term in a document, and a global weight, g_i, which describes the relative frequency of the term within the entire collection of documents.

Some common local weighting functions are defined in the following table.

Binary	L_{ii} =1 if the term exists in the document, or else 0
Term frequency	$L_{ii}=tf_{ij}$ the number of occurrences of term i in document j
Log	$L_{ij}=\log(tf_{ij}+1)$
Augnorm	$Lij=(tf_{ij}/max_i(tf_{ij})+1)/2$

Some common global weighting functions are defined in the following table.

Binary	$G_i=1$
Normal	$Gi=1/\sqrt{\sum_j (tf_{ij}^2)}$
Gfidf	$Gi=gf_i/df_i$ where gf_i is the total number of times term I occurs in the whole collection, and df_i is the number of documents in which term i occurs.
Idf	$Gi=\log_2(n/1+df_i)$
Entropy	$Gi=1+\sum_j (p_{ij}\log p_{ij}/\log n)$, where $p_{ij} = tf_{ij}/gf_i$

Empirical studies with LSI report that the Log and Entropy weighting functions work well, in practice, with many data sets. In other words, each entry a_{ij} of A is computed as: $a_{ij} = g_i \log(tf_{ij}+1)$.

Evaluation parameters for the IRs

The performance of identifying correct documents by the IRs which are indexed by some of the indexing techniques according to the user query has been measured via three metrics: precision, recall, and F-measure. Precision is the percentage of correctly identified documents over all the documents indexed by the IRs, while Recall is the percentage of correctly identified document over all the correctly identified document and unidentified document. Suppose the number of correctly identified web document is C, the number of wrongly identified web documents are W and the number of not identified web document is M, then the precision of the approach is given by the expression given below

$$P = C/ (C + W) \quad (1)$$

and the recall, R, of the approach is

$$R = C/ (C + M) \quad (2)$$

F-measure incorporates both precision and recall. F-measure is given by

$$F = 2PR/ (P + R) \quad (3)$$

3. NEED OF INDEXING IN IRS

i. This system is adopted by search engine to retrieve information in the form of documents according to the user query.

Ii.Grouping of documents within a cluster and class, according to keyword based or semantic based.

iii. Various page rank algorithms are used in IRs, which assigns wattage to the web documents, which further used by search engine to give ranked result to the user according to its query.

iv. Information discovery, automated document classification, Text summarization,

Relationship discovery, Automatic generation of link charts of individuals and organizations, Matching technical papers and

grants with reviewers, Online customer support, Determining document authorship, Automatic keyword annotation of images Understanding software source code.

4. CONCLUSION

At last we make a conclusion that, indexing in information retrieval is a process of finding and fetching the knowledge based information from cluster or collection of documents. This REVIEW paper deals with the basics of the information retrieval. In first section we are defining the information retrieval system with their basic measurements and evaluation parameter of a good IRs. After this we concerns with traditional IR models and also discuss about the indexing technique used in Boolean retrieval model of IRs. This paper also includes the applications of IRs.

5. 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 Techniques",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.