Performance Based Study and Comparative Analysis of Various Algorithms and Techniques for Information Retrieval

Jinal H. Tailor
Asst. Professor, Gujarat Technical University, Ahmedabad
Email: jinal.tailorssa@gmail.com
Gaurang K. Panwala
Asst. Professor, Gujarat Technical University, Ahmedabad
Email: gaurangpanwala.2013@gmail.com

--ABSTRACT---

The Technique and process of Storing. Searching and recovering Information is the science of searching and transfer of information in form of metadata from documents or searching within databases, whether relational stand-alone databases or hypertext networked databases such as the Internet or World Wide Web or intranets, for text, sound, images or data. The IR is mainly concern with retrieving relevant documents to a query to fulfill user's need as well as efficiently retrieving required data from large set of documents. The basic overview of some algorithmic problems arising in the representation of text/image/multimedia objects in a form amenable to automated searching, and in conducting these searches efficiently. These operations are central to information retrieval and digital library systems. The basic aim of different algorithm analysis is fast and efficient retrieval of user required data from large data set. While searching based upon different objects such as image, sound, documents process is depends upon different IR matrix techniques.

Keywords - **Clustering, Corpus, Digital Library ,Morphological, Stemmer**

I. INTRODUCTION

IR provides extracted set of data in various forms depending upon user's requirements. It describes set of data that is suitable for processing. IR system efficiently represent the filtered information i.e. output of query.

Information retrieval (IR) systems were originally developed to help manage the huge scientific literature that has developed since the 1940s. Many university, corporate, and public libraries now use IR systems to provide access to books, journals, and other documents. Commercial IR systems offer databases containing millions of documents in myriad subject areas. Dictionary and encyclopedia databases are now widely available for PCs. IR has been found useful in such disparate areas as office automation and software engineering. Indeed, any discipline that relies on documents to do its work could potentially use and benefit from IR. Different algorithm analysis mechanism provides efficient and effective way to search through the large amount of data from Web.

II. APPLICATION AREAS OF IR

- Digital library : one kind of library that stores information in digital form and accessed by computers
- Recommender system : Recommender systems or recommendation engines form or work from a specific type of information filtering system technique that attempts to recommend information items (films, television, video on demand, music, books, news, images, web pages, etc) that are likely to be of interest to the user. Typically, a recommender system compares a user profile to some reference characteristics, and seeks to predict the 'rating' that a user would give to an item they had not yet considered. These characteristics may be from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).
- Search engines: A search engine is one of the most the practical applications of information retrieval techniques to large scale text collections. Web search engines are best known examples, but many others searches exist, like: Desktop search, Enterprise search, Federated search, Mobile search, and Social search. Relevance feedback is an important issue of information retrieval found in web searching. Reliability of information is a pre-requisite to get most from research information found onto the web.
- Media search: An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common methods of

image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation.

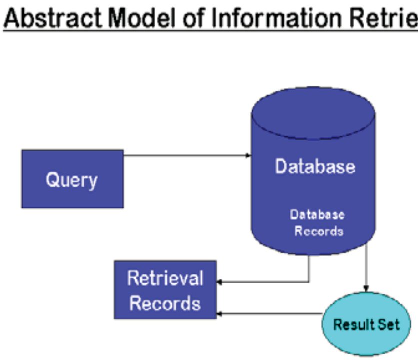


FIG 1

III. IR ALGORITHMS

We can distinguish three categories of algorithms for the IR system. Other algorithms are inherited from other computer applications. These include three different classes such as retrieval indexing and filtering algorithms.

Retrieval Algorithms

The main class of algorithms in IR is retrieval algorithms, that is, to extract information from a textual database. We can distinguish two types of retrieval algorithms, according to how much extra memory we need: Sequential scanning of the text: extra memory is in the worst case a function of the query size, and not of the database size. On the other hand, the running time is at least proportional to the size of the text, for example, string searching. Indexed text: an "index" of the text is available, and can be used to speed up the search. The index size is usually proportional to the database size, and the search time is sub linear on the size of the text, for example, inverted files and signature files . Formally, we can describe a generic searching problem as follows: Given a string t (the text), a regular expression q (the query), and information (optionally) obtained by preprocessing the pattern and/or the text, the problem consists of finding whether $t\epsilon\sum^*q\epsilon\sum^*$ (q for short) and obtaining some or all of the following information:

1. The location where an occurrence (or specifically the first, the longest, etc.) of q exists. Formally, if $t\epsilon\sum^*q\epsilon\sum^*$, find a position m >=0 such that $t\epsilon\sum$(from 0 to m)$q\epsilon\sum^*$. For example, the first occurrence is defined as the least m that fulfills this condition.

2. The number of occurrences of the pattern in the text. Formally, the number of all possible values of m in the previous category.

3. All the locations where the pattern occurs (the set of all possible values of m). In general, the complexities of these problems are different. The efficiency of retrieval algorithms is very important, because we expect them to solve on-line queries with a short answer time. This need has triggered the implementation of retrieval algorithms in many different ways: by hardware, by parallel machines, and so on.

Filtering Algorithms

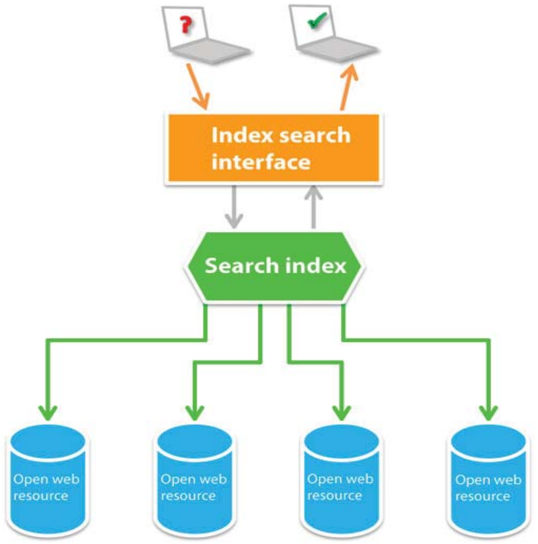
This class of algorithms is such that the text is the input and a processed or filtered version of the text is the output. This is a typical transformation in IR, for example to reduce the size of a text, and/or standardize it to simplify searching. The most common filtering/processing operations are :

•Common words removed using a list of stop words;
•Uppercase letters transformed to lowercase letters;
•Special symbols removed and sequences of multiple s
Paces reduced to one space;
•Numbers and dates transformed to a standard format;
•Word stemming (removing suffixes and/or prefixes);
•Automatic keyword extraction;
•Word ranking. Unfortunately, these filtering operations may also have some disadvantages. Any query,
Before consulting the database, must be filtered as is the text; and, it is not possible to
Search for common words, special symbols, or uppercase letters, nor to distinguish text fragments that have been mapped to the same internal form.

Indexing Algorithms

The usual meaning of indexing is to build a data structure that will allow quick searching of the text. There are many classes of indices, based on different retrieval approaches. For example, we have inverted files, signature files, tries, and so on. Almost all type of indices is based on some kind of tree or hashing. Perhaps the main exceptions are clustered data structures (this kind of indexing is called clustering), which is covered in further laboratories, and the Direct Acyclic Word Graph (DAWG) of the text, which represents all possible sub words of the text using a linear amount of space and is based on finite automata theory. Usually, before indexing, the text is filtered. Figure 2 shows the complete process for the text.

The preprocessing time needed to build the index is amortized by using it in searches. For example, if building the index requires O (n log n) time, we would expect to query the database at least O (n) times to amortize the preprocessing cost. In that case, we add O (log n) preprocessing time to the total query time (that may also be logarithmic).



Vector Space Models

Many of the text document search operations are conveniently performed using the vector-space representation of documents. The added advantage of this representation is that it can be extended to more general objects that include, for instance, images and video. Representing multimedia objects. A major advantage of the vector-space representation is that it is not specific to text documents; in fact it is used in virtually all current multimedia retrieval systems. The main idea is that instead of terms, we now represent the presence/absence of a set of features that are extracted from the object. In an image one might, for instance, record for each primary color the average density of that color; each color might become an axis in the space.

Algebraic Methods

Given the vector space representation of objects, it is natural to consider the use of algebraic methods to facilitate retrieval. An intriguing technique known as Latent Semantic Indexing does just this. We revert for a moment to the discussion of text retrieval, although all our comments here can be generalized to vector space representations of arbitrary objects. The main idea is that vectors representing the documents are projected down to a new, low-dimensional space obtained by the singular value decomposition of the term-document matrix A. This low-dimensional space is spanned by the Eigen vectors of ATA that correspond to the few largest eigenvalues and thus, presumably, to the few most striking correlations between terms. Queries are also projected and processed in this low-dimensional space.

IV. ALGORITHM ANALYSIS

One approach to improving precision in text retrieval is to use categorized search, each document is assigned to one or more categories. Typically, the categories are described by descriptive names; a user may then restrict their search to one of the categories. In basic search engine operation main categories are divided into different sub categories. One issue that arises is that in hierarchical categories, the importance of a term for searching depends on the position in the hierarchy. For instance, the term “computer” is useful in deciding whether a document should lie in the category “computer science”; within that category, though, it is essentially useless at distinguishing sub-categories and can be considered a stop-word. How should the categories be chosen so as to improve the relevance of the documents returned by the search? To being with, the categories should be chosen in a manner that is intuitive to the anticipated user-population (rather than to the person designing the system). Next, the categories should be reasonably balanced: a categorization in which a small set of categories contains most of the documents is not likely to be as useful as a balanced taxonomy. Finally, the categories should “span” the corpus in the following sense: it is not very useful to have two categories that are very similar to each other (this actually makes it harder for the user to decide which category to select when searching). A typical algorithmic operation in building and maintaining taxonomies is clustering. The idea is to partition the document into clusters of closely-related documents One consequence is that the categorization need no longer be static (as above); instead, clustering partitions the documents into related clusters each of which may be described, perhaps, by the terms most frequently occurring in the cluster. (This may not always be the most intuitive representation from the standpoint of the user, but preliminary experience with the technique is encouraging. Further, the clustering can change, as new documents are added (and old ones deleted); thus in a corpus of news articles, the clustering may change as the focus of the news changes. Such dynamic clustering, again, could be hierarchical.

V.CONCLUSION

IR becomes more complicated task to perform because of multiple combination phrases used by different users. New availability of hypermedia makes it more difficult to search desired document by simple keyword or the index or by hierarchical order of indexing or hyperlink through sequence of related documents.

One technique for improving IR performance is to provide searchers with ways of finding morphological variants of search terms. If, for example, a searcher enters the term *stemming* as part of a query, it is likely that he or she will also be interested in such variants as *stemmed* and *stem*. We use the term *conflation*, meaning the act of fusing or combining, as the general term for the process of matching morphological term variants. Conflation can be either manual--using some kind of regular expressions--or automatic, via programs called *stemmers*. Stemming is also used in IR to reduce the size of index files. Since a single stem typically corresponds to several full terms, by storing stems instead of terms, compression factors of over 50 percent can be achieved.

REFERENCES

[1] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using Linear algebra for intelligent information retrieval. SIAM Review, 37(4), 1995, 573-595, 1995.

[2] D. R. Cutting, J. 0. Pedersen, D. R. Karger and J. W. Tukey. Scatter/Gather: A Cluster-based Approach To Browsing Large Document Collections. Proceedings Of ACM SIGIR. 318-329. 1992.

[3] K. Jarvelin and J. Kekalainen. IR evaluation methods for Retrieving highly relevant documents. In SIGIR 23, pages 41–48, 2000.

[4] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, May 1999.

[5] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: two sides of the same coin? Communications of the ACM, 35(12):pp. 29–38, December 1992.

[7] Mehrdad Dianati, Insop Song, and Mark Treiber. An introduction to genetic Algorithms and evolution strategies. Technical report, University of Waterloo, Ontario, N2L 3G1, Canada, July 2002.

[8] P. Vakkari and N Hakala. Changes in relevance criteria and problem stages in task performance. Journal of Documentation, 5(56):540562, 2000.

Biographies and Photographs

First Author: Jinal H. Tailor

Currently working as Asst Professor at Computer Science Dept, Gujarat Technical University, Ahmedabad. She has completed her Bachelors in Computer Application with Distinction from Veer Narmad South Gujarat University, Surat. She did Master in Computer Application with Distinction from Nirma University, Ahmedabad. Her Area of Interests is Information Retrieval, Data Mining, Green Computing, Virtual World, Data Encryption.

Second Author: Gaurang K. Panwala

Currently working as Asst. Professor at Computer Science Dept, Gujarat Technical University, Ahmedabad. He has completed his Bachelors in Computer Application with Distinction from Veer Narmad South Gujarat University, Surat. He has done Masters in Computer Application with First Class from Gujarat Technical University, Ahmedabad. His areas of Interest are Information Representation, Clustering, Statistical Analysis, Data Structure and Networking Concepts.