SYNOPSIS

TEXT MINING FOR INFORMATION RETRIEVAL

Introduction

Nowadays, large quantity of data is being accumulated in the data repository. Usually there is a huge gap from the stored data to the knowledge that could be constructed from the data. This transition won't occur automatically, that's where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. Fast developing computer science and engineering techniques and methodology generates new demands to mine complex data types. A number of Data Mining techniques (such as association, clustering, classification) are developed to mine this vast amount of data. Previous studies [18] on Data Mining focus on structured data, such as relational and transactional data. However, in reality, a substantial portion of the available information is stored in text databases (or document databases), which consists of large collections of documents from various sources, such as news articles, books, digital libraries and Web pages. Text databases are rapidly growing due to the increasing amount of information available in electronic forms, such as electronic publications, e-mail, CD-ROMs, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database).

Data stored in text databases is mostly semi-structured, i.e., it is neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, length, category, and so on, but also contain some largely unstructured text components, such as abstract and contents. In recent database research, studies have been done to model and implement semi-structured data. Information Retrieval techniques, such as text indexing, have been developed to handle the unstructured documents. But, traditional Information Retrieval
techniques become inadequate for the increasingly vast amount of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual or user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, Text Mining has become an increasingly popular and essential theme in Data Mining.

Text Mining, also known as knowledge discovery from text, and document information mining, refers to the process of extracting interesting patterns from very large text corpus for the purposes of discovering knowledge. It is an interdisciplinary field involving Information Retrieval, Text Understanding, Information Extraction, Clustering, Categorization, Topic Tracking, Concept Linkage, Computational Linguistics, Visualization, Database Technology, Machine Learning, and Data Mining [25].

Text Mining tools/applications intend to capture relationship between the data. They can be roughly organized into two groups. One group focuses on document exploration functions to organize documents based on their content and provide an environment for a user to navigate and browse in a document or concept space. It includes Clustering, Visualization, and Navigation. The other group focuses on text analysis functions to analyze the content of the documents and discover relationships between concepts or entities described in the documents. They are mainly based on natural language processing techniques, including Information Retrieval, Information Extraction, Text Categorization, and Summarization [27], [28].

Content-based text selection techniques have been extensively evaluated in the context of Information Retrieval. Every approach to text selection has four basic components:

- Some technique for representing the documents
- Some technique for representing the information needed (i.e., profile construction)
- Some way of comparing the profiles with the document representation
- Some way of using the results of the comparison
Issues in Information Retrieval

The main motivation of our research is to study different existing tools and techniques of Text Mining for Information Retrieval (IR). Search engine is the most well known Information Retrieval tool. Application of Text Mining techniques to Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query.

Electronic information on Web is a useful resource for users to obtain a variety of information. The process of manually compiling text pages according to a user's needs and preferences and into actionable reports is very labor intensive, and is greatly amplified when it needs to be updated frequently. Updates to what has been collected often require a repeated searching, filtering previously retrieved text web pages and re-organizing them. To harness this information, various search engines and Text Mining techniques have been developed to gather and organize the web pages. Retrieving relevant text pages on a topic from a large page collection is a challenging task.

Given below are some issues identified in Information Retrieval process:

Issue (1): Traditional Information Retrieval techniques become inadequate to handle large text databases containing high volume of text documents. To search relevant documents from the large document collection, a vocabulary is used which map each term given in the search query to the address of the corresponding inverted file; the inverted files are then read from the disk; and are merged, taking the intersection of the sets of documents for AND, OR, NOT operations [8],[24],[30],[31]. To support retrieval process, inverted file require several additional structures such as document frequency of each lexicon in the vocabulary, term frequency of each term in the document. The principal cost of searching process are the space requirement in memory to hold inverted file entries, and the time spend to process large size inverted files maintaining record of each document of the corpus as they are potential answers. Many terms in the query means more disk accesses into the inverted file, and more time spent to merge the obtained lists.
Issue (2): Presently, while doing query based searching, search engines return a set of web pages containing both relevant and non relevant pages, sometimes showing non relevant pages assigned higher rank score. These search engines use one of the following approaches to organize, search and analyze information on the web. In the first approach [30], ranking algorithm uses term frequency to select the terms of the page, for indexing a web page (after filtering out common or meaningless words). In the second approach [5],[9],[11],[19],[20] structure of links appearing between pages is considered to identify pages that are often referenced by other pages. Analyzing the density, direction and clustering of links, such method is capable of identifying the pages that are likely to contain valuable information. Another approach [9],[26],[29] analyzes the content of the pages linked to or from the page of interest. They analyze the similarity of the word usage at different link distance from the page of interest and demonstrate that structure of words used by the linked pages enables more efficient indexing and searching. Anchor text [15] of a hyperlink is considered to describe its target page and so target pages can be replaced by their corresponding anchor text. But the nature of the Web search environment is such that the retrieval approaches based on single sources of evidence, suffer from weaknesses that can hurt the retrieval performance. For example, content-based Information Retrieval approach does not consider link information of the page while ranking the target page and hence affect the quality of web documents, while link-based approaches [6],[19],[20] can suffer from incomplete or noisy link topology. The inadequacy of singular Web Information Retrieval approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous retrieval strategy for Web Information Retrieval.

Issue (3): A common problem of Information Retrieval is that users have to browse large number of documents containing both relevant and non relevant documents before finding relevant documents. Clustering keeps similar documents together in a single group and hence fastens the process of Information Retrieval by retrieving the documents from the same cluster based on query vector matching to the cluster centroid. There are many clustering algorithms available like K-means, Bisecting K-means, HFTC (Hierarchical Document clustering using Frequent Itemsets), Hybrid PSO+K-means method and Global K-means [3],[7],[16],[17],[23]. But, there are many challenges in using these existing clustering techniques in the domain of text documents. Bisecting K-means produce deep hierarchy resulting in difficulty to browse
if one makes an incorrect selection while navigating a hierarchy. Although HFTC [4] produces a relatively flat hierarchy as compared to Bisecting K-means, but it is expensive in terms of calculating the global frequent itemsets to create the clusters. Global K-means [22], unlike K-means, is insensitive to the choice of initial $k$ cluster centers, thus giving global optimum solution. But it requires execution of K-means method $(nk)$ times for document set of size $n$ to generate $k$ clusters showing time complexity of $O(nk)$. In Hybrid PSO+K-means method [12], the PSO (Particle Swarm Optimization) module is executed for a short period to search for the optimum cluster centroid locations and then the K-means module is used for refining and generating the final optimal clustering solution. This method produces the global optimum solution like Global k-means, and also requires assuming the initial value of $k$.

**Issue (4):** To fasten the process of document retrieval, text summarization technique is used [1],[2],[10],[13]. Ranking of documents is made based on the summary or the abstract provided by the authors of the document. But it is not always possible as not all documents come with an abstract or summary. Also when different summarization tools like Copernic, SweSum, Extractor, MSWord, Intelligent, Brevity, Pertinence text summarizer are used to summarize the document, not all the topics covered within the document are reflected in its summary.

**Proposed Solutions**

The aim of this thesis is to address the above discussed issues in order to improve the accuracy of the Information Retrieval process. Our proposed approaches in this thesis contribute in the field of text Information Retrieval and provide more relevant documents (web pages) to the given searched query accurately and efficiently in less time.

1) To speed up the process of text document retrieval and effective utilization of the memory space as discussed in issue (1), we propose an algorithm based on inverted index file. By using the range partition feature of oracle, the space requirement of memory is reduced considerably as the inverted index file is stored on secondary storage and only the required portion of the inverted index file is maintained in the main memory. Fuzzy logic is applied to retrieve the selected documents and then suffix tree clustering is used to group the similar documents.
2) To handle the problem discussed in issue (2), a method is proposed for learning web structure to classify web documents and demonstrates the usefulness of considering the text content information of backward links and forward links for computing the page rank score. The similarity of the word usage at single level link distance from the target page is analyzed, which shows that content of words in the linked pages enables more efficient indexing and searching. The novel method efficiently reduces the limitations of the already existing Link Analysis algorithms like Kleinberg’s HITS algorithm, SALSA [14], [21] while computing the rank score of the retrieved web pages and the results obtained by the proposed method are not biased towards in-degree or out-degree of the target page. Also the rank scores obtained showing non-zero values help to rank the web pages more accurately.

3) An approach to text document clustering that overcomes the drawback of K-means and Global k-means as discussed in issue (3) is proposed which gives global optimal solution with time complexity of $O(lk)$ to obtain $k$ clusters from an initial set of $l$ starting clusters.

4) A new method of building the generic, extract based single text document summary of fixed length is proposed to handle the limitations of existing summarization algorithms as discussed in issue (4). Index term(s) of the document is/are identified based on keyterms of each sentence and paragraph within the document. Rank of the sentence is computed based on the number of matching terms between the document and sentence index terms. Sentences having high rank score are extracted to be included in the final summary.

### Results and Findings of the Work done

Various conclusions and findings derived on applying the proposed methods to handle the above discussed issues in Information Retrieval are discussed below:

1) Compression and partitioning of inverted file reduces the memory space requirement and store the entire compressed inverted file in secondary storage. It is the use of compression and partitioning that results in the superior performance of retrieval process as shown in fig. 1. Through the range partitioning feature of Oracle, a smaller, faster representation for sorted lexicon of interest can be achieved. To process a query, only a small portion of the compressed inverted index file is cached in memory. Input/output (I/O) time required for loading a much smaller compressed postings list is
small, although it adds some cost of decompression. So, the retrieval system runs faster on compressed posting lists than on uncompressed posting lists. Also, sorted lexicon permits rapid binary search for matching strings. Conjunctive queries are easily handled through the concept of fuzzy logic in retrieving the documents having high value of $\alpha$-cut (threshold value) in document set for AND operation and non-zero value of $\alpha$-cut for OR operation. The proposed method also retrieves the documents having the synonyms of the searched query terms instead of query term itself, in the document.

Fig. 1 Compressed inverted index file structure showing range partition

2) On the basis of the results obtained, it is found that both the text content information of backward and forward links is useful for ranking the target page. It is also observed that utilizing only extended anchor text from documents that link to the target document or while just considering the words and phrases on the target pages (full-text) does not yield very accurate results. We analyze the similarity of the word usage at single level link distance from the page of interest and demonstrate that content of words in the linked pages enables more efficient indexing and searching. The proposed method efficiently reduces the limitations of
the already existing Link Analysis algorithms (HITS, pSALSA, SALSA, HubAvg, AThresh, HThresh, FThresh, BFS) while computing the rank of the web page and the results obtained by the proposed method are not biased towards in-degree or out-degree of the target page as the links to the target page can be easily differentiated as navigational, functional and noisy links. While computing the rank score of the target page, only functional links are considered. Also the non-zero rank scores obtained by the proposed method help to accurately rank the web pages while other Link Analysis algorithms sometimes compute zero rank score of the pages.

3) The following K-means and Global K-means clustering weaknesses are removed on applying the proposed clustering method:

(i) The number of clusters $k$, need not be assumed initially as required in K-means. This number $k$ is determined by the proposed clustering method itself. The required number of clusters are then obtained iteratively by combining the similar clusters based on their inter cluster distance (between the two clusters centroids) and minimum intra cluster distance (between the cluster centroid and its corresponding member documents).

(ii) Proposed method always obtain the same clusters, using the same data, even if the documents are considered in different sequence, which is not possible in K-means.

(iii) Different initial condition (k cluster centers) produces same cluster results. Hence the algorithm is not trapped in the local optimum as in K-means.

(iv) In the proposed method, it is not required to know which term in the document contributes more to the grouping process since we assume the TF-IDF weight of each term to determine their importance in the clustering process. Hence the final clusters produced are independent of any initial assumptions needed, at the start of the clustering process.

(v) The time complexity of the proposed clustering method is $O(lk)$ starting with $l$ initial clusters, which is less than the time complexity of Global K-means which is $O(nk)$ for a document set of size $n$, at the cost of cluster quality. But the gain from the reduce time complexity overrules this slight degradation in quality of clusters.
Experimental evaluation on Reuters newsfeeds (Reuters-21578) shows clustering results (entropy, purity, F-measure) obtained by proposed method comparable with K-means and Global k-means.

4) It has been observed that the proposed method to generate generic, extract single text document summary, clearly depicts the topic(s) discussed in the text document and shows linking between the sentences of the summary. Unlike other summarization methods, the method is independent of the structure of text document and the position of sentence within the document. A sentence appearing later in the document can be included in the summary according to its importance within the paragraph of the document. Our proposed text summarizer avoids redundant information in the summary by excluding the sentences conveying same information and hence improves the quality by including more information to the fixed length generated summary. We evaluated our approach on DUC-2002 corpus (dataset contain two baselines 100 words extract summary) and it shows satisfactory results when compared to all the reported summarization systems in terms of ROUGE-N (N=1 to 8).

Chapter details of the Thesis

The thesis is organized chapter wise as follows:

Chapter 1: This chapter is devoted to introduction about Data Mining, Text Mining and Information Retrieval. Different techniques, applications areas and architecture of Data Mining and Text Mining are discussed in the chapter. Basic concepts, models and techniques of Information Retrieval such as extraction of index terms, retrieval models are also discussed. At the end of the chapter, the different Information Retrieval evaluation techniques and the framework of Information Retrieval are explained.

Chapter 2: In this chapter, a discussion on related work on document indexing, hyperlink structure of web pages, clustering and text document summarization is discussed. Based on the literature survey on each topic, the problems and challenges identified from existing tools and techniques for each are discussed in brief, providing the basis for the work to be carried out.
Chapter 3: This chapter discusses the method for *Quick Text Retrieval Algorithm Supporting Synonyms Based on Fuzzy Logic*. In this chapter, different compression algorithms (like Elias Gamma code, Elias Delta code, Fibonacci Code) are studied to store inverted index file and the concept of Fuzzy Information Retrieval is discussed along with Suffix tree clustering.

Chapter 4: This chapter is about *Web Page Ranking Based on Text Content of Linked Pages*. In this chapter, different link analysis ranking algorithms (HITS, pSALSA, SALSA, HubAvg, AThresh, HThresh, FThresh, BFS) are discussed and the ranking scores of pages computed through these link analysis ranking algorithms are compared with the proposed ranking approach which compute the rank score of the target web page based on the content analysis of the link pages of the target page.

Chapter 5: It discusses the problem of *Automatic generation of initial value K to apply K-means method for Text Documents Clustering*. Different clustering techniques and their limitations are discussed like k-means clustering include Bisecting K-means, HFTC clustering (Hierarchical Document clustering using Frequent Itemsets), Hybrid PSO+K-means method, Global K-means method. A method of clustering is then proposed to overcome the limitations of the already existing clustering methods.

Chapter 6: This chapter contains two sections. In first section, *Document Summarization based on Sentence Ranking Using Vector Space Model* is discussed. In this section, different summarization tools are analyzed like Copernic, SweSum, Extractor, MSWord, Intelligent, Brevity, Pertinence. The summary obtained from these tools are obtained and compared with the proposed summarizer on DUC-2002 dataset using ROUGE package. In second section, a method is suggested to obtain *query-based text summarization using clustering* (both for single and multi-document).

Chapter 7: It is the last chapter of the thesis in which *conclusion and future scope* have been discussed.
**Keywords:** Information Retrieval, Suffix-tree clustering, Fuzzy logic, Query processing, Vector Space Model, Index compression, Inverted Index file, Backward links, Forward links, Link Structure Analysis, Web page ranking, Clustering, K-means clustering, Global K-means clustering, Extract Summary, ROUGE tool, Text Summarization

**List of Author’s Publications**


**Major References**


Synopsis - 13


