



Improving Text Categorization Using Configuration File

V. Jayaraj¹ and V. Mahalakshmi²

¹School of Computer Science, Engineering & Applications, Bharathidasan University, Tamilnadu, India.

²School of Computer Science, Engineering & Applications, Bharathidasan University, Tamilnadu, India.

Abstract: This paper describes the brief history of the research and development of information retrieval systems. This paper opens with the review of an early development in information retrieval system. Information Retrieval system is concerned with effective and efficient retrieval of information from the unstructured text. With the growth of the unstructured information around the world increases and it leads to the need of an information retrieval system. An IR system searches the relevant information in the collection of unstructured or Semi structured data and it retrieves the needful information for an user's query.

Keywords: Structured Data, Unstructured Data, Indexing, Query Expansion.

I. INTRODUCTION

In this computerized world, as everything being computerized, many new techniques and technologies have been emerged. One of the latest commonly used technique is named as Mining. Data Mining is the process of extracting the information from large amount of data, stored in the database. In database, the data are stored in an ordered manner. Likely, another mining technique named Text Mining is the process of extracting the text from an unordered textual database. The text mining is also known as Text Data Mining. To accomplish this, countless techniques have been emerged and utilized. A technique that helps business intelligence is data mining. But data mining has some restrictions as it can be efficient in mining the information from the structured internal data and predicting the trends based on these data or knowledge to make wise decision. But some of the business handles the textual information which is not in a structured manner such as project report, employee resume, and profile [6]. These documents are presented in unstructured form, since the details are prescribed in the form of text using natural language and each of the documents followed the own developers style. Unlike the data in the database, the information in the textual document are disperse through the text. From such kind of documents, the information can be mined as per the user wish and it can be implemented by the algorithm after developing a configuration file to identify the useful information from the document, and various patterns can be easily extracted and organized in order to provide the meaningful information. Through the extracted information, the administrator can able to make the decision to enhance the business.

It is very difficult for the user to retrieve the data from the database. To solve this problem many techniques have been implemented and still require enhancements to overcome many retrieval problems. Based upon the user need, a new idea has been addressed, while creating an xml file based on configuration it improves efficiency while retrieving the data. A substantial portion of the information are stored in text databases. Text databases are growing due to the increasing amount of unstructured information in the news articles, www, email etc. In traditional information retrieval, many available documents will be relevant to an user query so the user needs different tools to compare the documents, so many text mining tools have been used to capture the relationship between the data. Information Overloading, rise the amount of information and makes it difficult for researchers to find relevant information. During search operation, Queries and data should be processed by search engines and response be sent to information consumers. Information retrieval using text mining is tedious process. Since the textual data may of different categories and also the requirements of user may be different conditions. Mining the text from such different categories based upon different conditions set by an user may become tedious process, when the textual consists of more number of pages. In existing techniques, mining from such textual data based upon different conditions by reading the text line by line may reduce the performance and increase the time. But, we overcome this drawback by proposing the methodology to improve the performance by providing proper response for the user requirements. Information Extraction framework is designed, and it achieves better results than other models and it is feasible to classify the data.

II. ISSUES IN INFORMATION RETRIEVAL SYSTEM

Traditional Information Retrieval system become inadequate to handle large text databases, to search a relevant document from the large document collection, a vocabulary is used which map each term given in the search query. To support the retrieval process, inverted file need an additional structures such as document frequency of each lexicon in the vocabulary. During query based searching search engine returns both relevant and non-relevant pages, and it uses one of the following approach to organize, search and analyze information on the web. First it uses the ranking algorithm and then identifying the pages that are often pass on by other pages and also affects the quality of the web documents. Relevant and Non-relevant documents are present while browsing the documents. Cluster keeps similar documents and retrieves and hence the process of information retrieval based on query vector matching.

Information Retrieval system aims to retrieve the needed information from an unstructured data. While retrieving the relevant information many kind of issue occur in an retrieval system. Information retrieval systems have tended to solve it with different features. Many powerful approaches has also been used. One of the main techniques is indexing [1]. Such an system can provide more accurate, relevant search results even to short queries. Due to dealing with the unstructured information in the internet, many problems are facing by users. Peoples do not always submit the correct queries. Moreover, Searchers will yield high quality results even if the query terminology is imprecise. By using the information retrieval system, an query expansion will yield an good results. A continues problem occur in an IR system is how to deal with multiple terms that refer to the same concept [7]. In Query expansion a search term will given us input, looks for synonyms in the controlled vocabulary and it returns the document that matches the search term [2]. A reciprocation to this problem, explains the way to index documents so that the query will relent the high quality search results.

III. DISCUSSION

The intend of this work is to provide more relevant documents to the given searched query accurately in less time. To speed up the process of text document retrieval, template based file concept is used, the file can able to retrieve the data efficiently. Text Summarization can also avoid redundant information in the summary by excluding the sentence conveying same information and improve the quality. if the user wants to mine the text from the information stored in a word document or an excel file or in any text or pdf file [3]. In our methodology, we have to create a configuration file based upon the user requirements on extracting the text. In that configuration file, we have to set up a condition to extract the kind of text from the textual database. With the help of the configuration file, we have to train the data to mine the text from an unordered textual database [4]. Through this paper, we have to mine the text as per user wish from any pattern of the file.

Information retrieval using text mining is a tedious process, since the textual data may of different categories and also the requirements of the user may be different conditions. Mining the text from such different categories based upon different conditions set by the user may become tedious process, when the textual data consists of more number of pages. In existing techniques, mining from such textual data based upon different conditions by reading the text line by line may reduce the performance and increase the time. But, we overcome this drawback by proposing the methodology information retrieval configuration file to improve the performance by providing proper response for the user requirements. Information retrieval configuration file achieves better result. The proposed methodology has been experimentally verified and the result proves that the IRCF methodology satisfies the aim of the paper. To do the experiment, we have to undertaken 100 candidate's resume, which is of different format. From those 100 candidates, a particular group of members has to be selected. First, the selection process can be carried out using the existing algorithm namely "*KNN Text Classification Algorithm*" by examining each resume and identifying the qualifying candidate. The execution time and the scan time are noted and compared with that of the proposed algorithm. Next, the selection process can be carried out by using the proposed IRCF text mining algorithm.

No. of Resumes	Execution Time in (mins)		Scan Time in (Sec)	
	KNN	IRCF	KNN	IRCF
10	10	2	96	53
20	25	5	120	92
50	45	7	230	156
75	65	10	313	190
100	90	12	465	212

Table 1: Result of time optimization

Based on the algorithm, the qualifying candidate can be chosen and the time period to obtain the result has to be noted. By comparing these two time periods, we can conclude that the proposed algorithm performs well than the existing methods in table 1 and also comparison chart shows in fig 1 and fig 2. Thus our proposed Information retrieval configuration file algorithm satisfies the user need and helps the decision makers to improve the business intelligence. Thus our proposed methodology satisfies the aim of the paper and helps the decision makers to improve the business intelligence.

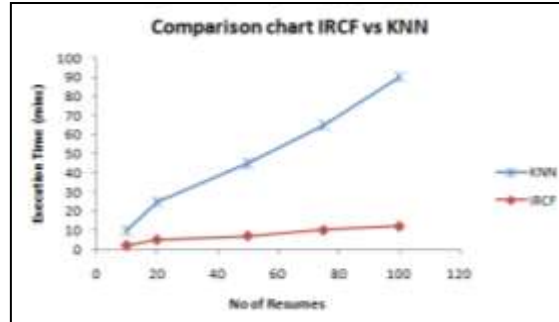


Fig. 2: Comparison chart during execution time

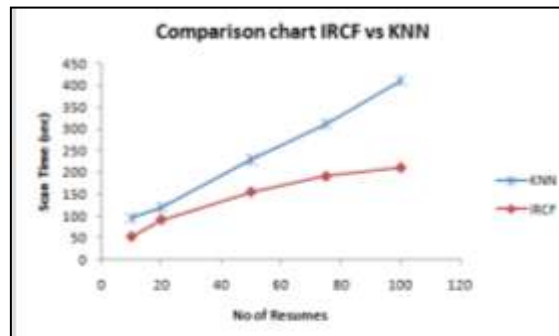


Fig. 3: Comparison chart during scan time

V. CONCLUSION

Information Retrieval using configuration file extract the relevant document when compared to the other algorithm during retrieval. Many Researches are developing an algorithm for improving the scope of business intelligence. The extraction process is completed successfully and the text from unstructured document is converted into a structured table. From the value stored in the table, the decision maker can make the decision to improve their process. The future enhancement of this research work is to implement the xml file and to optimized the data retrieving time.

VI. REFERENCES

- [1]. Diana Inkpen . Information Retrieval on the Internet, University of Toronto.
- [2]. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
- [3]. V.Jayaraj, V.Mahalakshmi. IRCF Text Mining Algorithm for Text Categorization for Improving Business Intelligence(2014). International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS) . ISSN (Online): 2279-0055
- [4]. V.Jayaraj, V.Mahalakshmi. An Empirical Study of Extracting information for Business Intelligence (2014). International Journal of Engineering, Business and Enterprises Applications (IJEBA). ISSN (Online): 2279-0039.
- [5]. Yang, Y., Pedersen J. O. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (1997) 412–420.
- [6]. Kun Yu, Gang Guan, Ming Zhou. Resume Information Extraction with Cascaded Hybrid Model, Proceedings of the 43rd Annual Meeting of the ACL, June 2005.
- [7]. Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering, Vol. 4, No. 2, April, 2009.