Efficient Searching Of Research Paper Using Hierarchical Clustering

[1]Mr. Satish Krishna Aurange, [2]Prof. S. B. Siledar

[1,2]CSE Dept, MITCOE, Aurangabad, Maharashtra, India.

[1]*satish.aurange@gmail.com*

Abstract - The time spent by users are almost two or more hours looking for papers that generates the possibility to make a search engine to optimize and precision in the results. This works purposes a better classification of research papers, the architecture works with a database of knowledge. That's the initial work of a classification using text mining techniques to search into the documents with natural language contained and get the best words of their content to get a database knowledge, that's the first step to get the desired knowledge also the proposed work use the some engine to make searches classifying the information introduced by the final user and searching in the correct cluster.

Keywords — text mining, cluster means , database ,pattern, knowledge.

I. INTRODUCTION

On the internet user spend lots of hours for searching one of the information. Something will also happen when researcher wants to work on some particular topic or researcher want to make some research work on some of the topic. Researcher has to spend the lots of hours on some website or some search engine then. Researcher has to find some the research paper by using some advanced search option. When researcher want to search the research paper for information retrieval then this will be done by using some advanced searching with Google or any one of the search engine[1].By using above said technique the research paper will be found and this paper, some-time, may not appropriate. Because of this, time spent by user is wasted. Much research has been carried out about the search for similar words in textual, mostly for applications in information retrieval tasks. The basic assumption of most of these approaches is that words are similar if they are used in the same contexts. The methods differ in the way the contexts are defined (the document, a textual window, or more or less elaborate grammatical contexts) and the way the similarity function is computed.

II. LITERATURE SURVEY

The large hours spent by user makes necessary to develop a prototype to enable analysis of the performance for testing based in the amount of accurate results related to the type of search performed as well as the relationship obtained in articles. Since the obtained knowledge base must be taken as a starting point to determinate patterns within a word captured and to deduce the weight to be given by the user submit this information to the data mining algorithm. The problem of document clustering is generally defined as follows: Given a set of documents, would like to partition them into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic and the documents in different clusters represent different topics. In most existing document clustering algorithms, documents are represented using the vector space model which treats a document as a bag of words [4].

The main problem is to solve the next points:

- Develop architecture for the management and use of knowledge base adapting a correct interpretation of patterns related to each sentence.
- Implement an architecture using a ms access and create database knowledge
- Develop a filter to generate a classification of research papers and also searches
- Develop a wrapper to generate a classification of research papers and also searches
- The time searching papers takes between two and four hours to locate the correct paper.

III. OBJECTIVE

The main objective is to implementation of data mining to solve a problem related to searching a research paper .The other objectives are as follows.

- To minimize the searching time
- To get appropriate research paper
- To get paper classified according to their topic

DOI : 10.18231/2454-9150.2018.0030

- To get paper which is correct and appropriate for the better result
- To give the list of papers which is according to frequency of words found in the research paper

IV. System Architecture

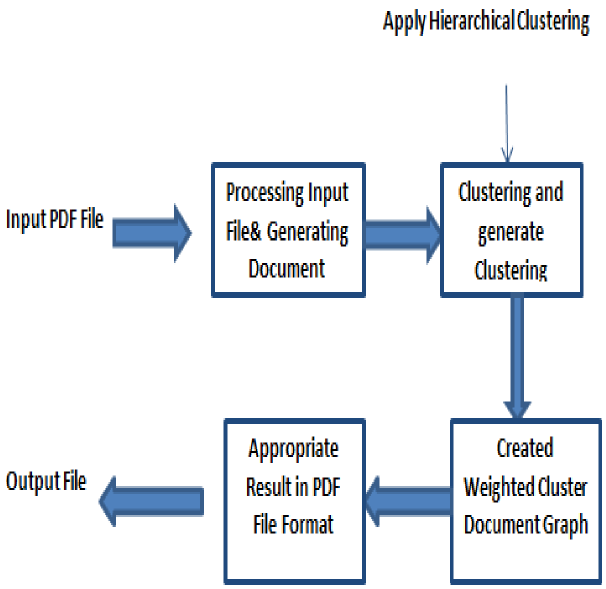
Apply Hierarchical Clustering
Input PDF File
Processing Input File& Generating Document
Clustering and generate Clustering
Output File
Appropriate Result in PDF File Format
Created Weighted Cluster Document Graph

Fig.1. System Block diagram

The total workflow is divided into following modules

1. Processing input file and generating document graph

This block is needed to accept the file. It is responsible to upload file, to process the file i.e. to form nodes for every newline contents. It is also responsible for generating weight from each node to very other node.

2. Clustering node and building clustered graph

This block is responsible for choosing a clustering algorithm out of two. It also accepts the threshold, so that can check the similarity between the clusters up to that level. It is responsible for making clusters.

3. Creating weighted document clustered graph

This block is responsible to accept the fired query. It is responsible to check the similarities between the query a contents and the contents in the clusters. It then build weighted clustered document graph.

4. PDF generation

This block is responsible for generating the PDF which contains the highest number of word (query) which was gave as input from the clusters we formed, as a response for fired query. It generated the minimal clusters and after finding the weight of the node for fired query.

V. PROBLEM OUTLINE

The users spend a lot of hours searching in the repositories of papers on topics related to the area of information technology and development, which requires the establishment of a search engine to locate items of research[5] in the area of programming languages allowing identification of basic patterns in the input text and the implementation of a data mining algorithm to help decrease the response time in the search within the database for locating scientific articles and as a prototype-level implementation that allows access from a browser. The main problem is to solve the next points:

- Develop a Filter to generate a classification of research papers and also searches
- Develop a Wrapper to generate a classification of research papers and also searches
- The time searching papers takes between two and four hours to locate the correct paper. The search engine must have a better time than the actual.

VI. RESULT ANALYSIS

Time Comparison

Table 1.1 shows the comparison of time required for existing system and proposed system. Perform the operation on three data set and compare the results. It shows the time required in ms for compendium system and feature based system.

Table 1.1 Time Comparison Table

Search Text	Existing System	Proposed System
P1	124	62
P2	3578	2813
P3	2124	2124
P4	2844	2750
P5	2719	2375
P6	2954	2718

Figure 1.1 shows the comparison of time required for existing system and proposed system. In the graph X-axis shows the number of research paper and y Axis shows time in ms.:

Memory Required Results

Table 1.2 shows the comparison of memory required for existing system and proposed system. Perform the operation on three dataset and compare the results. It shows the memory required in bytes for existing system and proposed system.

Table 4.3 Memory Comparison Table.

Research Papers	Existing System	Proposed System
P1	4452	708
P2	35900	28255
P3	30297	22482
P4	33762	23555
P5	32429	27396
P6	48083	38050

DOI : 10.18231/2454-9150.2018.0030

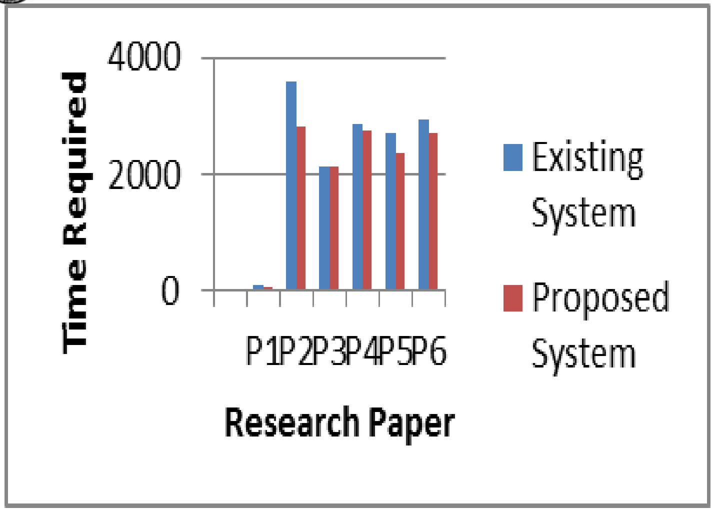

Figure 1.1 Execution time Graph.

Figure 1.2 shows the comparison of memory required for existing system and proposed system. In the graph X-axis shows the number of research paper and y axis shows memory requires in bytes.

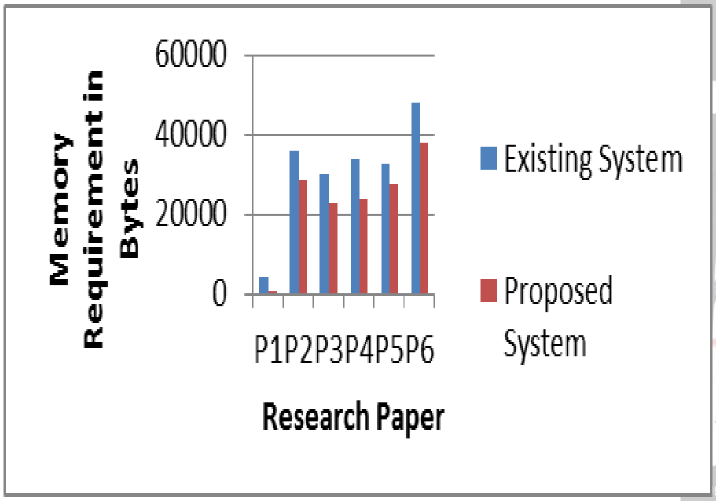

Figure 1.2 Graph for Memory Comparison.

VII. CONCLUSION

Implemented system to efficient categorize the research paper based on hierarchical clustering .The proposed system shows the following improvement,

Precision has been improved by 23.29 Recall has been improved by 15.96 F-Measure has been improved with difference of 0.9

In addition to the above improvement the proposed system has been performed checked on the amount of compression, execution time taken and memory required .All these parameter have shown significant improvement of 0.14, 250.16 and 7412.83 respectively with respect to k-means clustering method.

It can be concluded that hierarchical clustering is capable for efficient searching of research paper.

REFERENCES

[1] E. Alan Calvillo,Alejandro Padilla,Jaime Munoz "Searching Research Paper Using Clustering and Text Mining" 2013

[2] K. Suresh, D.K., S. Ghosh , S. Das and Ajith A., "Automatic Clustering with Multi-Objective Differential Evolution Algorithms".

[3] "A Brief Survey of Text Mining". Andreas Hotho KDE Group University of Kassel Andreas N¨urnberger Information Retrieval Group School of Computer Science May 13, 2005.

[4] M. Mete, X. Xu, Chun-Yang F., Gal Shafirstein, "Head and Neck Cancer Detection in Histopathological Slide, International Workshop on Data Mining in Bioinformatics", Sixth IEEE International Conference of Data Mining (ICDM 2006), December 18-22, 2006, Hong Kong.

[5] Szymanski, J., Self-Organizing Map Representation for Clustering Wikipedia Search Results. 2011.

[6] "Integrated Clustering and Feature Selection Scheme for Text Documents" Journal of Computer Science 6 (5): 536-541, 2010. ISSN 1549-3636 © 2010 Science Publications

[7] "Text Mining: The state of the art and the challenges". Ah-Hwee Tan Kent Ridge Digital Labs 21 HengMuiKeng Terrace Singapore 119613

[8] "Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK" (International Journal of Computer Science & Security (IJCSS), Volume (4) : Issue (2) 176)

[9] "Streaming Hierarchical Clustering for Concept Mining" http://www.arl.wustl.edu/ projects/fpx/ reconfig.html).

DOI : 10.18231/2454-9150.2018.0030