Search Engine Using Clustering and Text Mining

Vinod S. Badgujar[1], Asha H. Pawar[2]

[1] *Student, Computer Science & Engineering* ,ZES's Dnyanganga college of Engg. & Research, *Maharashtra, India*
[2] *Assist Prof, Computer Science & Engineering* ,ZES's Dnyanganga college of Engg. & Research, *Maharashtra, India*

---***---

Abstract - ***The time spent by users are almost two or more hours looking for papers that produces the chance to make a search engine to enhance and accuracy in the results. The proposed work is to organize research papers, using a database of knowledge related with the topics of programming, databases and operating systems. Using Clustering technique the database is created for the required search. There are numerous clustering algorithms such as hierarchical clustering, self-organizing maps, K-means clustering and so on. In this paper, we propose a clustering algorithm that search into the documents with natural language contained and get the best words of their content to form a database knowledge that the first step to get the desired knowledge. We implemented the system using the K-means clustering algorithm. Moreover the future work uses the search engine to make searches classify the information introduced by the last user and searching in the exact cluster.***

Key Words: *Search Engine , Knowledge Base, Key Text Mining, Mining.*

1. INTRODUCTION

The use of search engines to locate information has grown steadily based on the needs of users generating a snowball effect, including information that is not useful or significant but also added data of scientific interest to remember that not all the information is by default [1]. The generation of multiple research papers goes to a generation of multiple repositories. This kind of segmentation of information makes a generation of hours of search papers when a researcher it's searching for a specific topic. The main goal to minimize the time overturned in searches and also makes best searches and has a minimal time of search [2]. But searching normally, only responsible for presenting results on screen as a search interface running on multiple search engines. The implementation of informative search engines has been evolving as the needs of the research sector. The clustering is an unsupervised classification of patterns into groups the clustering problem has been addressed in many contexts and by result in many disciplines; this reflects its broad appeal as one of the steps in exploratory data analysis. However, clustering is difficult problem combinatory and differences in expect and contexts. [3]and differences in assumptions. The implementation of a better tool to search research articles which would be useful to the result and minimize times of search and make a best engine to get best results in every search of research papers by considering not only the title but the contents also. The use of K-Means algorithm allow us to implement semi-supervised learning clusters using an algorithm so as to help identify approximate the text to search using predefined patterns and the implementation of a cluster algorithm [4] for consultations within the database manager MySQL (database manager that allows free use of multithreading, multi-search and multi-user) in order to obtain scientific research papers.

This work is organized as track in section 2 the paper give an introduction of the problem statement presented in the search of research papers also in section 3 a solution to the problem outline with architecture to classify and locate research papers. In section 4 gives implementation of study where show how the architecture works in a user environment. Finally, section five shows a target of related papers

1.1 PROBLEM STATEMENT

The users spend a lot of time searching in the repositories of papers on topics related to the area of interest for the research, 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 text mining algorithm to help decrease the response time in the search within the database(MySQL) for locating required articles and as a prototype-level implementation. A lot of time 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 and the relationship obtained in articles or papers. 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 text mining algorithm. The main problem is to solve the next points:

- Of knowledge base adapting a correct Interpretation of patterns related to each sentence.

- Implement an architecture using a MySQL server and create database knowledge
- Develop a Filter to produce a classification of research papers and searches
- Develop a Wrapper to generate a classification of research papers and searches
- The search engine must have a better time than the actual.

2. PROPOSED WORK

The implementation of data mining to solve a problem involves the need to implement a methodology focus into the analysis of pattern into the texts, where there are several methodologies custom built-oriented type of attributes that will be reviewed, such methodologies are not recommended for our implementation as the proposed work need a methodology to be adaptive evolutionary behavior.

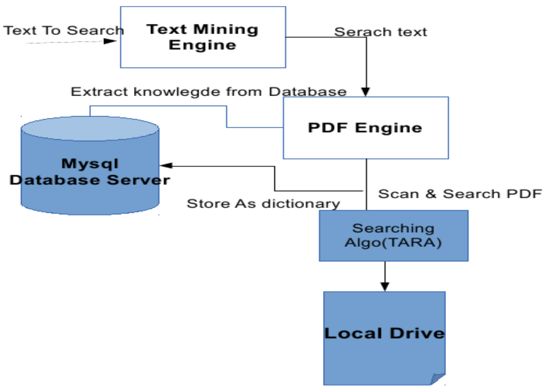


Fig -1: Search Architecture Model

In figure 1 shows the search architecture model, the main purpose is the use of text mining and also the primary part of the architecture needed in our problem where the user begins a search interface for entering text within the information used in the process of searching patterns within a knowledge base in order to obtain parameters for the selection of cluster where the search was implemented, once achieved the search is conducted within the database in the process of localization papers. This current architecture works using as input the location of the research to get text patterns in pdf, that work reading line by line and in this process is supported by a knowledge base that is fed into a first semi-automatically with the information collected from items previously stored.

This System Architecture is divided into two parts:

*1. **Search Architecture:***

The important use of text mining and the first part of the architecture wanted in our difficulty where the user begins a search interface for entering text within the information used in the process of searching patterns within a knowledge base in order to obtain parameters for the selection of cluster where the search was implemented [1] once achieved the search is conducted within the database in the process of localization papers.

Algorithm of search patterns:

1. Opening PDF file
2. Read PDF line
3. Compare the contents of the line with the information that is in the knowledge base
4. Comparing whether there is a similarity of at least 80% between the line and the value of the knowledge base.
5. According to the result the Cluster is assigned
6. Return to step 2 until the file ends.

*2. **Pattern matching Architecture**:*

We search the relevant pattern from the research paper to search the knowledge pattern from the database, with this the proposed work generate our engine to make pattern matching needed in the problem outline start with a pattern matching that read the article searching a similar pattern compared with the knowledge base once they locate in which it relates is selected cluster on which will be uploaded from the article in the database [1].

The use of clusters involves the classification of different groups partitioned that share a characteristic in this way determines a measure of characteristics between the stored information in the knowledge database [1].

3. K-MEANS CLUSTERING

The implement of clusters will be using the K-Means algorithm which will be used to send the parameters for classification of research papers in this case the search engine will use five clusters to achieve implementation [6]. The algorithm works by using the following equation

$$\arg\min \sum_{i=1}^{k} \sum_{x_j \in s_i} || Xj - \mu i ||2 \quad \text{........[1]}$$

The formula represents a given set of observations (X1, X2… Xn) where each observation represent an element of the cluster with a d-dimensional real vector[1] , k-means clustering aims to partition and the n observations into k sets (k<= n) S = { S1, S2, ….. Sk) that's to minimize the cluster where μi is the mean of points in Si [1].

***K-Means algorithm**:*

Step 1: Select objects randomly. These objects represent initial group centroids k.
Step 2: Assign each object to the group that has the closest centroids.
Step 3: When all objects have been assigned, recalculate the positions of the centroids k
Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

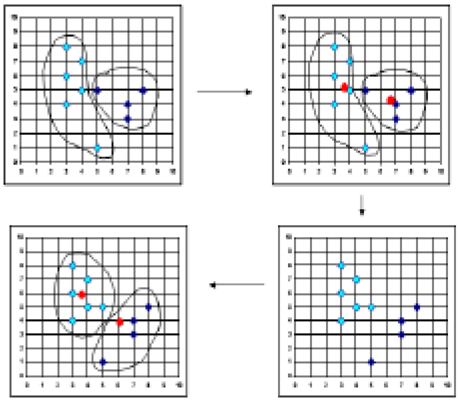
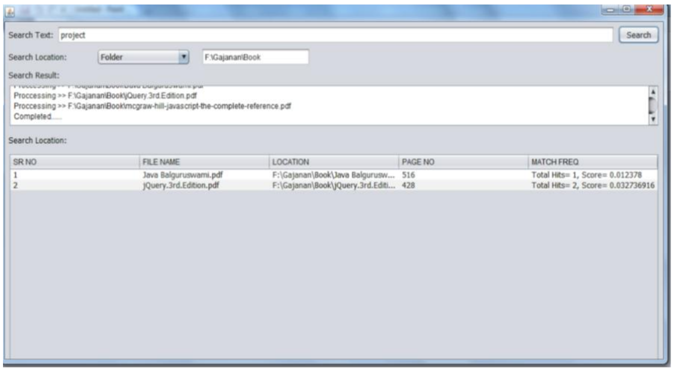


Fig.2 . K-Means Example

The centroids are calculated in k-means algorithm, arithmetic mean of the cluster all points of a cluster with the given distance measure distances are computed [8].

4. IMPLEMEANTATION

The implementation of the search engine using data mining scheme works by using a clustering where the user enter the information that want and finally showing a list of papers available. The Fig. 3 shows the interface to be used for searching of research papers with a simple structure that helps the user to identify a single text area where the user enter text on the search and finally at the bottom were generated the results of this search.



Search Text: project

Search Location: Folder F:\Gajanan\Book

Search Result:

```
Proccessing >> F:\Gajanan\Book\jQuery.3rd.Edition.pdf
Proccessing >> F:\Gajanan\Book\mcgraw-hill-javascript-the-complete-reference.pdf
Completed.....
```

Search Location:

SR NO	FILE NAME	LOCATION	PAGE NO	MATCH FREQ
1	Java Balguruswami.pdf	F:\Gajanan\Book\Java Balgurusw...	516	Total Hits= 1, Score= 0.012378
2	jQuery.3rd.Edition.pdf	F:\Gajanan\Book\jQuery.3rd.Editi...	428	Total Hits= 2, Score= 0.032736916

Fig 3. Interface of User

The time in the search takes one minute to search in all the database knowledge, making a best time than the usual when the researcher are searching. The search in the second column of results, with their own value of comparison using the full text in the search engine and compared with their own category in this case data base. Steps of K-means algorithm are

- The item is placed randomly in each cluster
- Compare the items without classification
- Items comparative review of their distance from each other using the mean of each element
- if it is near the item is added to the cluster, if not so return to step two
- Once the cycle are finished the elements clustered

Here first step assigns an item to each cluster at random to start with the clustering using the maintain provisions for the comparison and thus be systematically ordering the clusters which will make the search query [1].

5. CONCLUSIONS

This paper evaluates a way to optimize the information to be located within a structured framework with an initial Knowledge base. This helps the easy categorization of information by implementing a clustering for fast search and locations well as a textual analysis entered by the user as a basis for discussion, as future work is to implement an automatic learning which allows the steady increase in the manipulated texts.

This kind of techniques allows making the best search engine using database to work with filter, wrapper or even ontology. The uses of text mining technologies are not used in web search or meta search, that kind of tools usually use only meta crawler to classify the information the current work and shows how the search engine can be used and it should make a benchmark between the filter, wrapper and ontology to the next work.

ACKNOWLEDGEMENT

The authors thank the editors and reviewers for providing very helpful comments and suggestions. Their insight and comments will definitely led to a better presentation for the ideas expressed in this paper.

REFERENCES

[1] Searching Research Papers Using Clustering and Text Mining (978-1-4673-6155-2/13/ © 2013 IEEE).

[2] A Text Clustering System based on *k*-means Type Subspace Clustering and Ontology.(International Journal of Electrical and Computer Engineering 1:5 2006).

[3] K-means-like Algorithm for K-medoids and Its Performance, Department of Industrial and Management Engineering, POSTECH —In Proceedings. Of CCS "07, pp. 598–609, 2007.

[4] Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition.(Michael W. Berry and Malu Castellanos, Editors Jan 4, 2013).

[5] 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).

[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 Heng Mui Keng Terrace Singapore 119613)

[8] week 14 Data mining-Clustering-Classification-Wrap-up.