International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 12, December-2015

# Clustering of Web Search Results using Hybrid Algorithm

Anupreeta Mishra
Department of Computer Engineering,
MESCOE, SPPU,
Pune, Maharashtra, India 411007

Megha Prince
Department of Computer Engineering,
MESCOE, SPPU,
Pune, Maharashtra, India 411007

Suyog Tate
Department of Computer Engineering,
MESCOE, SPPU,
Pune, Maharashtra, India 411007

Nikhil Nirbhavane
Department of Computer Engineering,
MESCOE, SPPU,
Pune, Maharashtra, India 411007

Gopal Deshmukh
Department of Computer Engineering,
MESCOE, SPPU,
Pune, Maharashtra, India 411007

***Abstract*— Clustering the web search has become a very fascinating research area among scientific and academic associations involved in information retrieval. It is also knows as Web Clustering Engines, appeal to increase the description of documents presented to the user for review, while decreasing the time spent reviewing them. Many algorithms for web document clustering already exist, but conclusions show there is room for more algorithms. Our Project works on providing concise information on an ambiguous search. This allows the user to gain precise information faster and reduces the time spent on looking through thousands of pages for simple information. The information obtained will be segmented, sorted and irrelevant information will be avoided.**

***Keywords*— Data mining, web document clustering, at multiple granularities, web search engine.***

## I. INTRODUCTION

Our project focuses on retrieval of concise information using web search clustering in data mining. Data mining is a field in computer science which helps to determine hidden patterns and extract knowledge from large amount of data. Now a days, Data mining is used in business applications for precise decision making and getting relevant information from big data sets.

Clustering is a process of grouping set of similar objects such that they belong to same group and different from other groups. Clustering can be achieved by various methods or algorithms. There are various communities which use different methods of clustering for grouping of data. Clustering is normally used with data mining.

Text-mining techniques have been used for solving the problem by automatically classifying the text documents, mainly in English. A famous text-mining approach which is ontology-based used to cluster the research proposals based on their similarities in research fields. The technique is effective and efficient for clustering of research areas with both the languages whether it is English and Chinese. The method also includes an optimization model that considers applicants characteristics for balancing proposals by geographical regions. The results can also be used to improve the efficiency and effectiveness of research project selection processes in other government and private research funding agencies.

Many techniques have been introduced regarding this, namely, Semantic information, hierarchical, link based, iterative, suffix tree method.

Currently search engines like Google use iterative search to retrieve information. In this case the user receives large amount of data on an obscure search. Google, when searching sorts the list of links according to the number of times the page has been visited.

Our project uses Iterative Fuzzy C-means, which is a new description centric algorithm. Iterative Fuzzy C-means selects a maximum estimated number of clusters using Forgy's strategy, then it iteratively merges clusters till results cannot be improved automatically evaluating the best solution and number of clusters.

It also uses K-means which makes use of Euclidian distance formula to form clusters and in those clusters we use Iterative Fuzzy C-means to give weight according to each search.

## II. LITERATURE SURVEY

Optimal meta search results clustering: The rapid increasing of Web applications which are not based on longer texts represents both an challenge and opportunity to the mining algorithms which are text based, because of scattered representations and lack of distributed context [1]. To address this problem, investigation a term expansion approach based on analyzing the relationships between the term concepts present in the concept lattice associated with the document corpus. Definition of five term concept association measures: proximity, concept similarity, connection strength,

damping-weighted proximity, proximity and strength. By means of two case studies, evaluation of the effectiveness of these measures for expansion-enhanced K-NN classification and K-Means clustering of short texts. The results suggest that the five measures are highly competitive, with the best measure showing a clear improvement over the corresponding non enhanced K-Means and K-NN algorithms, as well as providing two alternative for expansion enhancements (i.e., based on Wordnet and on pseudo-relevance feedback)

A survey of Web clustering engines: As resources become more and more available on the Web, so the difficulties associated with finding the desired information increase. Intelligent agents can assist users in this task since they can search, filter and organize information on behalf of their users. The techniques of Web document clustering can help users to find pages which meet their requirements of information. TopicSearch is a personalized web document clustering algorithm [2]. To improve the query expansion, TopicSearch presents a new inverse document frequency function, a new algorithm which is memetic for web document clustering and recurrent phrases for defining cluster labels. Every user query is handled by an agent which it coordinates several tasks including query expansion, search results acquisition, pre-processing of search results, cluster construction and labelling, and visualization. The above tasks are performed by agents who are specialized whose parallelized execution in certain instances. The model was successfully tested on fifty DMOZ datasets. The results presented an improved precision and recall over traditional algorithms (STC y Lingo, Bisecting k-means, k-means). In addition, the model presented was assessed by a group of twenty users with over 90% in the favour of model.

Modern Information Retrieval: Addison Wesley Longman Publishing: Information retrieval (IR) has considerably changed in the last few years with the expansion of the World Wide Web and the introduction of inexpensive and modern graphical user interfaces and mass storage devices [3].

Due to this, Conventional IR textbooks have become fully out-of-date which has led to the launch of new IR books recently. Nevertheless, belief that there is still great need of a book that approaches the field in a rigorous and complete way from a computer-science perspective (in opposition to a user-centred perspective).

Subtopic retrieval methods: Clustering versus diversification of search results: To address the inability of ongoing ranking systems for supporting retrieval of subtopic, two important techniques for post-processing of search results have been examined: clustering and diversification [4]. By using a group of complementary evaluation measures which can be applied to both ranked lists and partitions and two test collections concentrating on ambiguous and broad queries, respectively. The chief finding of the experiments is that clustering is finer for full single subtopics retrieval, with a better balance is achieved in performance by generating numerous subsets of search results which are diverse while diversification is of top hits more helpful for faster coverage of subtopics which are distinct. There is scope for little improvement for the search engine baseline if not interested in strict retrieval of subtopics, and that search results clustering methods do not perform well on queries with low divergence subtopics, mainly due to the difficulty of generating discriminative cluster labels.

Mining: Clustering Web Documents A Preliminary Review: Web search result clustering of document has evolved as a good tool for improving performance of Information Retrieval (IR) system. Search results often troubled by problems like high volume, synonymy, polysemy etc [5]. Rather than solving these problems it also gives user the ease to find his/her required information. Here, WSRDC-CSCC a method is introduced to group (cluster) web search result by utilizing cuckoo search a meta-heuristic method and Consensus clustering. Cuckoo search provides a solid foundation for consensus clustering. As a local clustering function, k-means technique is used. The final number of cluster is not depended on this k. Consensus clustering finds the Natural grouping of the objects. The experimental results show that proposed algorithm finds the actual number of clusters with great value of precision, recall and F-measure as compared to the other method.

## III. MODULES

*A. Login: It will allow the user to login to the webpage.*

*B. Search: Normal searching*

*C. Search using C-means: searches the query by using the Fuzzy C-means algorithm*

*D. Web Usage Statistics: shows the search statistics of the user.*

*E. Logout: Terminates user Session.*

## IV. ARCHITECTURE

This project has the following components in its architecture User, Google server, Web Browser, database. Fig. 1 shows the system architecture.

### A. System Architecture

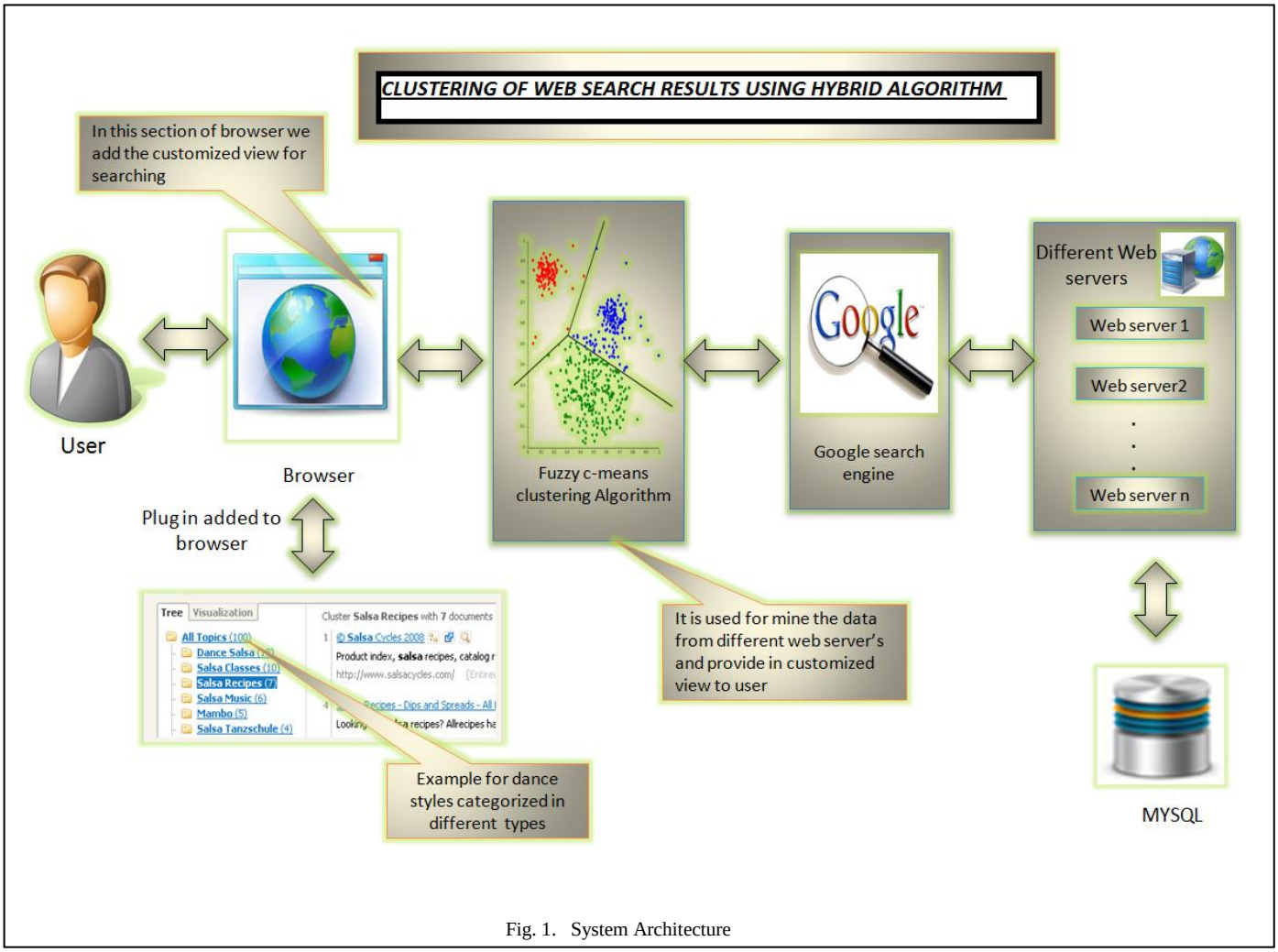


Fig. 1. System Architecture

*1) User Searches Query*

The query entered by user is pre-processed using the following:

*a) Apply NLP on Searched Query*

*b) Lower case filtering*

*c) Stop word removal*

*d) Porter's stemming algorithm*

*e) Term- Document matrix (TDM)*

*f) Elimination of dimensions with a range equal to zero*

*2) Browser connects to Google server*

After applying the above algorithm it will search on the Google for getting information from different web servers.

*3) Google Chooses nearest data centre to retrieve information*

The query provided is searched on Google for information retrieval and data

*4) Algorithms are applied*

Fuzzy c-means clustering algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the data point and cluster center. Closer the data is to the cluster center closer is its membership towards the respective cluster center. Clearly, summation of membership of each data point should be equal to one. After every repetition of algorithm, cluster centers and membership are updated according to the formula.

*5) The output is given to user*

Output is given to the user in descending order of weights of clusters.

## V. ADVANTAGES

There are a variety of search engines each having their own features. Internet search engines can help organize individual websites. They can also organize vast amount of information that can sometimes be scattered in various places. Search engines will also cut down irrelevant information generated.

## VI. CONCLUSION

This paper provides snapshot of completed, ongoing and emerging methods of clustering. The web search engine forwards the work such as finding the effective data in most informative manner and user interaction. The results will represent the data with only meaningful information thus saving time of retrieval for the user.

## VII. FUTURE SCOPE

The demand for the search engines by the IT professionals and many people from other disciplines or domains will continue to adopt new methods for the web search results.

*1) To improve the User Experience*

*2) To improve the way in which search results are Segmented and Displayed*

*3) Fast Information Retrieval*

## REFERENCES

[1] C. Carpineto and G. Romano, "Optimal meta search results clustering," presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, 2010.

[2] C. Carpineto, et al., "A survey of Web clustering engines," ACM Comput. Surv., vol. 41, pp. 1-38, 2009.

[3] R. Baeza-Yates, A. and B. Ribeiro-Neto, Modern Information Retrieval: Addison Wesley Longman Publishing Co., Inc., 1999.

[4] C. Carpineto, et al., "Evaluating subtopic retrieval methods: Clustering versus diversification of search results," Information Processing Management, vol. 48, pp. 358-373, 2012.

[5] K. Hammouda, "Web Mining: Clustering Web Documents A Preliminary Review," ed, 2001, pp. 1-13.