IOSR Journal of Computer Engineering (IOSR-JCE)
ISSN: 2278-0661, ISBN: 2278-8727, PP: 46-49
www.iosrjournals.org

Cluster Based Web Search

Prof.D.A.Nikam [1], Mr. Joshi Govind[2], Mr.Bhandari Nikhil[3]
Mr. Varma PramodKumar[4]

[1]*(Computer Science and Engg/Shivaji University, India)*
[2]*(Computer Science and Engg/Shivaji University, India)*
[3]*(Computer Science and Engg/Shivaji University, India)*
[4]*(Computer Science and Engg/Shivaji University, India)*

ABSTRACT : *Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose, one of them is Data Clustering. Clustering implies filtering results obtained from search engine and provide more flexible result In this paper Data Clustering is discussed along with various approaches and their analysis.*
Keywords - *clustering ,stemming ,stopword, filtering ,query* .

I. INTRODUCTION

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criteria for checking the similarity is implementation dependent. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. Precisely, data clustering is a technique in which the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses are to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

II. LITERATURE SURVEY

In Web People Search via Connection Analysis[1],the author has proposed that web search is difficult because it is hard for users to construct queries that are both sufficiently descriptive and sufficiently discriminating to find just the web pages that are relevant to the users search goal. Queries are often ambiguous: words and phrases are frequently polysemantic and user search goals are often narrower in scope than the queries used to express them. This ambiguity leads to search result sets containing distinct page groups that meet different user search goals. Often users must refine their search by modifying the query to filter out the irrelevant results. Users must understand the result set to refine queries effectively; but this is time consuming , if the result set is unorganized . Web page clustering is one approach for assisting users to both comprehend the result set and to refine the query. Web page clustering identifies semantically meaningful groups of web pages and presents these to the user as clusters. The clusters provide an overview of the contents of the result set and when a clusters selected the result set is refined to just the relevant pages in that cluster After clustering whatever the load of queries on a single machine which is treated as server will get distributed over the network using Support Vector Machine, which act as load classifier or distributor. Depending upon the capacities of machines each can handle the specific load &return unstable when it exceed the limit. It is useful to determine whether the machine is stable or not.

In Disambiguation Algorithm for People Search on the [2],the author has proposed that searching for entities, i.e., WebPages related to a person, location, organization or other types of entities is a common activity in internet search today. The clusters can be returned in a ranked order determined by aggregating the rank of the Web pages that constitute the cluster. With each cluster a summary description that is representative of the real person associated with that cluster is provided .The user can hone in on the cluster of interest to her and get all pages in that cluster, i.e., only the pages associated with required result There is significant interest in the problem of Entity Search, with several research efforts addressing this and related challenges. The motivation for that is the fact that Entity Search can provide a way to browse and analyze the returned information in a more structured way, ultimately enhancing web search capabilities and the user experience. In this paper disambiguation algorithm[2] is developed and then study its impact on People Search. The proposed algorithm

first uses extraction techniques to automatically extract significant' entities such as the names of other persons, organizations, and locations on each webpage. In addition, it extracts and parses HTML and Web related data on each webpage, such as hyperlinks and email addresses. The algorithm then views all this information in a unified way: as an Entity-Relationship Graph where entities are interconnected via relationships[4]. The algorithm gains its power by being able to analyze several types of information: attributes associated with the entities and, most importantly, direct and indirect interconnections that exist among entities in the ERgraph.

In Towards Breaking the Quality Curse A Web-Querying Approach to Web People Search[3],the author has proposed that, searching for people on the Web is one of the most common query types to the web search engines today. However, when a person name is queried, the returned WebPages often contain documents related to several distinct namesakes who have the queried name. The task of disambiguating[2] and finding the WebPages related to the specific person of interest is left to the user. Many Web People Search (WePS)[1] approaches have been developed recently that attempt to automate this disambiguation process[2]. Nevertheless, the disambiguation quality[2] of these techniques leaves a major room for improvement. This paper presents a new server-side WePS approach[1]. It is based on collecting co-occurrence information from the Web and thus it uses the Web as an external data source. A skyline-based classification technique is developed for classifying the collected co-occurrence information in order to make clustering decisions. The clustering technique[5] is specifically designed to (a) handle the dominance that exists in data and (b) to adapt to a given clustering quality measure. These properties allow the framework to get a major advantage in terms of result quality over all the latest WePS[1] techniques we are aware of, including all the 18 methods covered in the recent WePS[1] competition.

III. PROPOSED EXPERIMENTAL WORK

The application contains the following components.

- Web page retrieval for the query
- Pre-processing of WebPages
- Clustering

The working of all these components is carried out in pipelined manner. Fig 2.1 shows the system architecture which is described as follows:

- Web page retrieval for the query

In this module various features of document such as number of Hyperlinks, Image count, Parsing time, occurrences of specific word are determined to define category of document. These features are used by clustering algorithms[5] to create clusters[6] of documents.

- Preprocessing of web pages

Algorithms used for the preprocessing are Stemming algorithm for transforming text into grammatical root form, Stop Word Removal[5] to remove unwanted words, Lingo[5] to ensure that we can create a human perceivable cluster label and only then assign documents to it, & to extract frequent phrases from the input documents, hoping they are the most informative source of human readable topic descriptions.

E.g. of Stop Word Removal is as follows:- Input :- I saw a cat.
Output :- saw cat.
E.g. of Stemming is as follows :-
Input :- connected/connecting/connections
Output:- connect.

- Clustering

We used k-Means & Greedy k-Means algorithm for clustering .The K-Mean algorithm is useful for small dataset. In Greedy K-Means method, the groups are identified by a set of points that are called the cluster centers. The data points belong to the cluster whose center is closest. Existing algorithms for k-means clustering[5] are slow and do not scale to large number of data points and converge to different local minima based on the initializations Greedy K-Means algorithm overcomes this drawbacks and help users to find relevant documents easily.

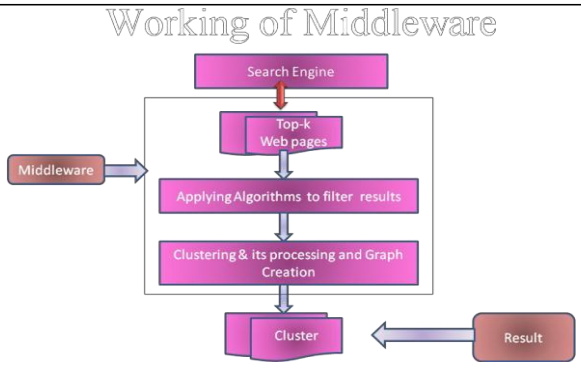

Fig 2.1 System Architecture

IV. TECHNIQUE TO BE USED

In this section various techniques for development of the application have been listed below

- Feature Extraction
- Pre-processing of extracted of web pages
- Clustering

Table 3.1 describes various techniques used for development of each module, which are listed as follows:

Table 3.1-Techiniques used

Module	Description	Technique Used
Feature Extraction	Various predefined features are extracted from the web pages in the dataset using Parser.	Parser
Preprocessing extracted pages	To provide a appropriate query by removing unwanted words to the clustering module	Stemming & Step Word
Clustering	Pre-process data above is passed as an input to the Greedy K-means Clustering algorithm which gives us well	Greedy K-means Algorithm

3.1.1 Feature Extraction*:* In this module various features of document such as number of Hyperlinks, Image count, Parsing time, occurrences of specific word are determined to define category of document. These features are used by clustering algorithms[5] to create clustersof documents. Working of this technique is shown in fig 3.1.1.

3.1.2 Preprocessing of pages retrieved*:* In this module keyword/query received from the user is appropriately processed so as it appear as proper input to the clustering technique. Working of this technique is shown in fig 3.1.1.

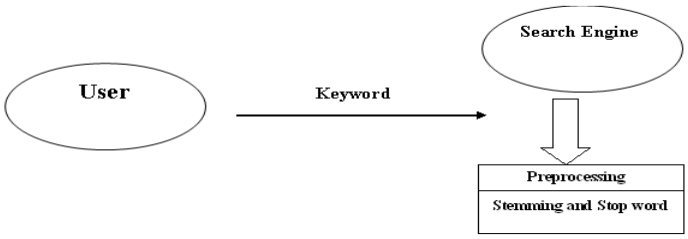

Fig 3.1.1:-Page Retrieving & Preprocessing

V. CONCLUSION

In this paper, we have tried to give the basic concept of clustering by first providing the definition .Then we have given different approaches to data clustering and also discussed some algorithms to implement that approaches. The use of some fast relevant algorithms is also possible which provides result in quicker time by more finer filtering which is possible in future time and is proposed future work.

VI. ACKNOWLEDGEMENT

We express our sincere thanks to Prof. D.A. Nikam whose supervision, inspiration and valuable guidance helped us a lot to complete the paper .Her guidance proved to be the most valuable to overcome all the hurdles in the fulfillment of this paper.

REFERENCES

[1] D.V. Kalashnikov, S.Mehrotra, R.N.Turen and Z.Chen, "Web People Search via Connection Analysis" IEEE
a. Transactions on Knowledge and data engg.Vol 20, No11, November 2008.0.
[2] D.V. Kalashnikov, S. Mehrotra, Z. Chen, R. Nuray-Turan, and N.Ashish, "Disambiguation Algorithm for
a. People Search on the Web," Proc. IEEE Int'l Conf. Data Eng. (ICDE '07), Apr. 2007.
[3] D.V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra, "Towards Breaking the Quality Curse. A Web-Querying
a. Approach to Web People Search," Proc. SIGIR, July 2008.
[4] R. Bekkerman, S. Zilberstein, and J. Allan, "Web Page Clustering Using Heuristic Search in the Web Graph," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 2007.
[5] 5) M. F. Porter. An algorithm for suffix stripping. Program Vol. 14, no. 3, pp 130-137.
[6] Data Clustering and Its ApplicationsRaza Ali (425), Usman Ghani (462), Aasim Saeed (464) .