CLUSTERING AND ORGANIZATION OF SEARCH RESULTS USING CLUSTERING ALGORITHM

Miss. Shefali A. Kedia[1], Prof. K. P. Wagh[2], Dr. Prashant N. Chatur[3]

[1]Department of Computer science and Engineering,

Government College of Engineering, Amravati, (India)

[2]Department of Information Technology,

Government College of Engineering, Amravati, (India)

[3]Department of Computer science and Engineering,

Government College of Engineering, Amravati, (India)

ABSTRACT

In recent years, internet has taken a very high pace and it has become the primary source of information. Any information about any object, person, place, etc. is first searched through web search engines. Even the smallest word's meaning is also searched on the web. Thus, it has reduced the manual efforts of people to collect information. The search results give a lot of results for the same query. We have to go through it to search for the most relevant result we are looking for. This problem is solved through the paper by giving clustered results for the ambiguous queries. We provide different sets of link for the same query thus reducing the effort of going through the results. We use k-means++ to cluster the links and then rearrange the results in colored format.

***Keywords:* Search-Results, Clustering, Rearrangement, Ambiguous.**

I. INTRODUCTION

The World is changing at a very high pace after the emergence of internet. The use of internet has become the daily chore of every individual across the world. From a small kid to an old aged person, every individual has a smart phone or tablet or a personal computer which is his primary object. Any information to be obtained abput anything is just a few seconds away. Internet has developed beyond our imaginations and can provide with almost all the details about all the things. It is the largest database of information currently.

Every information needed was first a little hard to get. People had to go through newspapers, books, articles, magazines, dictionaries, etc. to get the information. But, now-a-days it is just a button away. Everything is available on the internet. You just have to type and enter and everything is in your hands. Yes, we are talking about the search engines like Google, Yahoo, Bing, etc.

Any query to be resolved is just searched on the internet and within seconds all the information is obtained. But, a problem with this technology is that it has not fully reduced the work of an individual. It gives the highest visited results on the top in the form of a ranked list. Individual have to go through the results and get the best possible he is looking for. The most relevant result may vary from individual to individual. Like for example, a school going child may search "Apple" and expect the results of the apple fruit and a college going student may search "Apple" and expect the top results of the Apple Inc. Thus, the results of such ambiguous queries should be obtained in the clustered format and then the individual will find it easy to search is relevant domain. This paper attempts to get results in the form of clusters.

II. RELATED WORK

In this section, we discuss all the previous work done on the web search results and clustering.

1. TileBars:

M. A. Hearst [2], gave a searching technique add-on which works on Boolean queries. It gives the term distributed results at the beginning of each query result specifying the frequency of the words in document which are in the searched query.

2. Query Occurrence Visualization:

Heimonen and Jhaveri[3], tried to made an easy understanding of the obtained web results. They made the smaller icon of the web page of each link and added it to each link in search result page. This technique didn't help a lot.

3. Hotmap:

Orland Hoeber and Xue Dong Yang[4], used a heat scale to describe the results. The words with higher frequency has the darker shades of red and orange and the words with lesser frequency is in the lighter shades.

4. Tag Clouds:

Kuo, Hentrich B.M. Good and Wilkinson[5], introduced a new term known as tag clouds in which the link are represented in the form of tag clouds and their size, font and color change according to the frequencies.

5. Result maps:

Clarkson, Desai and Foley[6], stated a hierarchical technique to represent tag clouds in the form of a tree. The most ranked are kept above and least at the root.

6. Clustering on Similarity measure:

K. P. Wagh[9], proposed a method that uses similarity between the words and then cluster the results.

III. DESIGN LAYOUT

 ISBN: 978-93-86171-35-1

The traditional search techniques fail to provide an easy way to get to the relevant result for the ambiguous queries given by the user. Users have t go through each link to search the result he is looking for. This takes a lot of time. For e.g. the apple tree result after searching for “apple” is on the 8th page of google results. The starting pages all include the results of the Apple Inc. Thus, a person looking for information about the tree and not the company has to go to 80 links before he gets his desired result. Hence, a more easy approach has been proposed in this paper. It includes the following steps:

1. **Vector Representation:**
   A tf-idf vector of each snippet obtained is to be calculated . The stemming and stopword removal is done first and then the upper and lower cuts are established to calculate the tf-idf vector.
2. **Multi-dimensional Projection:**
   Each of the snippet is then projected on the 2-D graph by using multidimensional projection techniques.
3. **Clustering Algorithm:**
   A clustering algorithm K-means++ is used to cluster these snippets.
4. **Re-organization and Coloring:**
   Coloring different clusters with different colors and then re-organize it to get a comprehensive view.

IV. PROPOSED METHOD

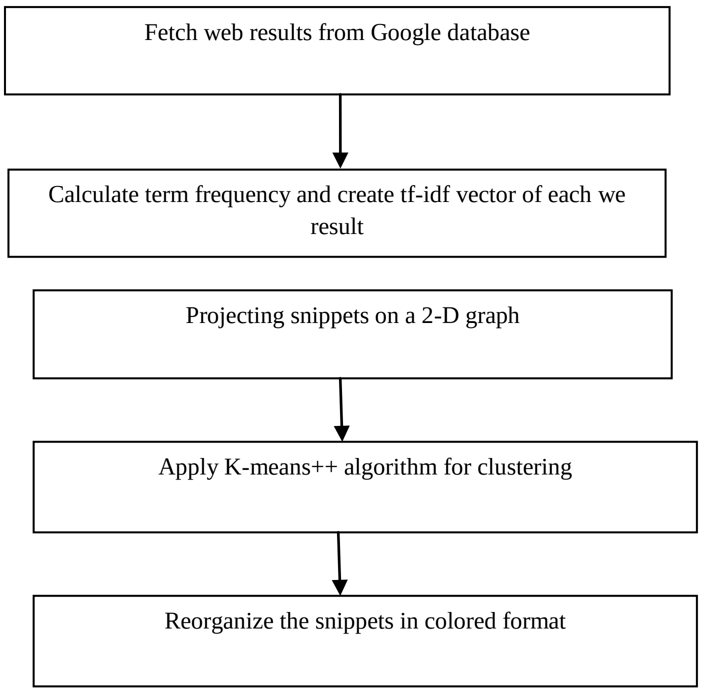

Fig. 1: System Architecture

Fig. 1 shows the architecture diagram of the system. It works as follows:

Step 1: The queries are fired through a search engine which fetches result from google search engine. Some top results are fetched from the Google and its snippets are extracted.

Step 2: Term frequency of each word of snippet is calculated and tf-idf vector of each is obtained.

Step 3: Now, each snippet is projected on a 2-D graph using some multi-dimensional technique.

Step 4: By applying k-means++ algorithm, snippets are clustered in clusters relevant to the query keyword.

Step 5: Snippets are then re-organized in colored form through their ranks.

Thus, this procedure provides a comprehensive view of the traditional search results.

V. CONCLUSIONS

In this paper, we stated a clustering method to organize snippet on the search result page to get a comprehensive view and easy to navigate through the results. It also lessens the time in which the results are obtained. The previous work done on it was complicated for the user to understand and the reorganization makes it user-friendly. The k-means++ algorithm helps it to be more efficient than other clustering methods.

REFERENCES

[1] Erick Gomez-Nieto, Frizzi San Roman, Paulo Pagliosa, Wallace Casaca, Elias S. Helou, Maria Cristina F. de Oliveira, “Similarity Preserving Snippet-Based Visualization of Web Search Results”, IEEE Transactions on Visualization and Computer Graphics, Vol. 20, pp. 457-463, 2014.

[2] M.A. Hearst, “TileBars: Visualization of Term Distribution Information in Full Text Information Access,” Proc. ACM SIGCHI Conf. Human Factors in Computing Systems, pp. 59-66, 1995.

[3] T. Heimonen and N. Jhaveri, “Visualizing Query Occurrence in Search Result Lists,” Proc. Ninth Int’l Conf. Information Visualisation, pp. 877-882, 2005.

[4] O. Hoeber and X.D. Yang, “The Visual Exploration of Web Search Results Using Hotmap,” Proc. 10th Int’l Conf. Information Visualization, pp. 157-165, 2006.

[5] B.Y.-L. Kuo, T. Hentrich, B.M. Good, and M.D. Wilkinson, “Tag Clouds for Summarizing Web Search Results,” Proc. 16th Int’l Conf. World Wide Web (WWW), pp. 1203-1204, 2007.

[6] E. Clarkson, K. Desai, and J. Foley, “Resultmaps: Visualization for Search Interfaces,” IEEE Trans. Visualization and Computer Graphics, vol. 15, no. 6, pp. 1057-1064, Nov. 2009.

[7] A. Spoerri, “RankSpiral: Toward Enhancing Search Results Visualization,” Proc. IEEE Symp. Information Visualization, pp. 208-214, 2004.

[8] M. Nizamee and M. Shojib, “Visualizing the Web Search Results with Web Search Visualization Using Scatter Plot,” Proc. IEEE Second Symp. Web Soc., pp. 5-10, 2010.

[9] PH Govardhan, K.P. Wagh, P. N. Chatur, “Web Document Clustering using Proposed Similarity Measure”, National Conference on Emerging Trends in Computer Technology,pp. 15-18, 2014.

[10] D. Arthur and S. Vassilvitskii, “k-Means++: The Advantages of Careful Seeding,” Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA), pp. 1027-1035, 2007.

[11] Sanjay D. Sawaitul, Prof. K. P. Wagh, Dr. P. N. Chatur, "Classification and Prediction of Future Weather by using Back Propagation Algorithm-An Approach", International Journal of Emerging Technology and Advanced Engineering, vol. 2, pp. 110-113, 2012.

[12] Rasika G.Charate, Dr.P.N.Chatur, Prof.K.P.Wagh, "Document Filtering: Intelligent Inference System for Web", International Journal of Management, IT and Engineering, vol. 3, pp. 207-222, 2013.

[13] S. Kolhe, K.P. Wagh, "Semantic Similarity based on Information Content", International Journal of Computer Science and Application Issue, pp. 82-86, 2010.

[14] S. Dehankar, K.P. Wagh, "Web Page Classification using Apriori Algorithm and Naive Bayes Classifier", vol. 3, April, 2015.