Clustering and Visualization on Web Search Results: A Survey

Shefali Kedia, Kishor Wagh and Prashant Chatur

Abstract A query fired on web search engines provide snippets in a ranked order of most visited once at the top. The results obtained are easy and understandable for the user. Certain queries those are ambiguous in nature fail to provide best match results. Clustering can solve this problem to a certain extent. The use of tf-idf vector followed by clustering through k-means++ and reorganization of the snippets will make user find the search more easy.

Keywords Web search result · K-means++ · Clustering

1 Introduction

Internet is taking over everything. A man's daily needs now include Internet. Any information needed is just a click away. Search engines play an important role in today's era of technology. Information required on any topic is searched and within seconds a lot of data is obtained. A query is fired and a ranked list of results also known as snippets is obtained. These snippets are short and can provide little information about the document inside. They are straightforward for queries that are unambiguous. The limitation is that a user has to explore the results to get desired domain. The results are ranked according to the most visited links. A user searching for fruit when types "apple" does not get the fruit link even in the first 100 links of result. Users then need to go through the links to find the desired result or give detailed query resulting in time consumption.

S. Kedia (✉) · K. Wagh · P. Chatur
Department of Computer Science and Engineering, Government College of Engineering, Amravati 444604, Maharashtra, India
e-mail: kediashefali@gmail.com

K. Wagh
e-mail: kishorwagh2000@gmail.com

P. Chatur
e-mail: prashant_chatur@rediffmail.com

P. K. Sa et al. (eds.), *Recent Findings in Intelligent Computing Techniques*,
Advances in Intelligent Systems and Computing 709,
https://doi.org/10.1007/978-981-10-8633-5_13

Clustering of web snippets [1] makes the searching for the user efficient and user friendly. The existing technologies improve snippets by using color glyphs, tag clouds, which add information about the documents.

Vector representation of snippets helps in knowing the word frequency of the unique words. The snippets are chosen as centers that are far away from each other. K-means++ is then applied to these snippets for clustering. After clusterization of these snippets, they are again made the tag clouds. This is done using different colors for different clusters that now becomes easy for the user to obtain desired result.

2 Related Work

2.1 Query Occurrence Visualization

Heimonen and Jhaveri [2] stated a scheme in which the query results are shown in a small icon which is placed beside the web search link. This scheme neither reduced time consumption of the task nor gave efficient results. Also, there was not any change in the number of users that use web search engines. Users said that it was unobtrusive and easy.

2.2 TileBars

Hearst [3] introduced a visualization technique TileBars, used for a Boolean query. It also provides the data on term distribution of the Boolean queries in the results. It gives the length of the web page, the term frequency, and query distribution. They are displayed on a tilebar column beside each search result. It helps in making the judgement about the web page easy.

2.3 Tag Clouds

Kuo et al. [4] replaced snippets by tag clouds and each tag's size, color, indexing is based on the snippet rank.

2.4 Hotmap

Hoeber and Yang [5] provide a heat scale to state the term frequency of the query term. The more frequency is given by red color and the lesser ones by the orange color. These are placed beside the snippet.

2.5 Rank Spiral

Spoerri [6] presented a rankspiral which scans large documents and the titles in a single window. It also uses spiral mapping that reduces occlusions and maximizes density. It solves the problem of labeling.

2.6 Resultmaps

Clarkson et al. [7] stated a hierarchical method that uses tree representation to represent the search results. The information is highlighted and indexed according to the ranks.

2.7 Clustering on Similarity Measure

Wagh [8] proposed a method of similarity measuring which measures the similarity of snippets to make clusters.

3 Design Rationale

Traditional list based results are not useful for ambiguous queries such as apple, jaguar, etc. or a broad topic. Visualization makes it more user friendly, easy to navigate the web search results. It could be an add-on on the traditional list-based technique than to be its replacement. It gives a more comprehensive view of the snippets.

3.1 Arrangement Based on Similarity

The tf-idf vector representation is made of each obtained snippet by calculating unique words and their frequency. The snippets are then grouped on similarity according to the keyword of query.

3.2 Ranking of Documents

The highest visited documents are ranked first in the clusters. Thus, visualization on web search results preserves the rankings of the results. The size of these link results can be resized and colored according to their ranks and clusters for better understanding.

3.3 Uncluttered Layout

The snippets those are similar to each other are grouped and may overlap when are projected on the graph. It can be difficult to find the relevant result the user is looking for. This can be solved by reorganizing using colors and ranks.

4 System Architecture

The steps of the proposed method are:

1. Preprocessing the obtained snippets and creating tf-idf vector for each snippet.
2. Plotting the vectors on graph and applying k-means++ algorithm to form clusters.
3. Organization and visualization of clusters according to the ranking by grouping it in different colors to get a comprehensive view and easy to understand (Fig. 1).

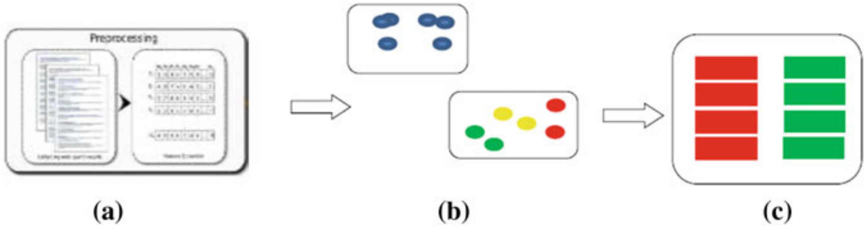

Fig. 1 **a** Preprocessing query. **b** Clustering snippets. **c** Organization of results

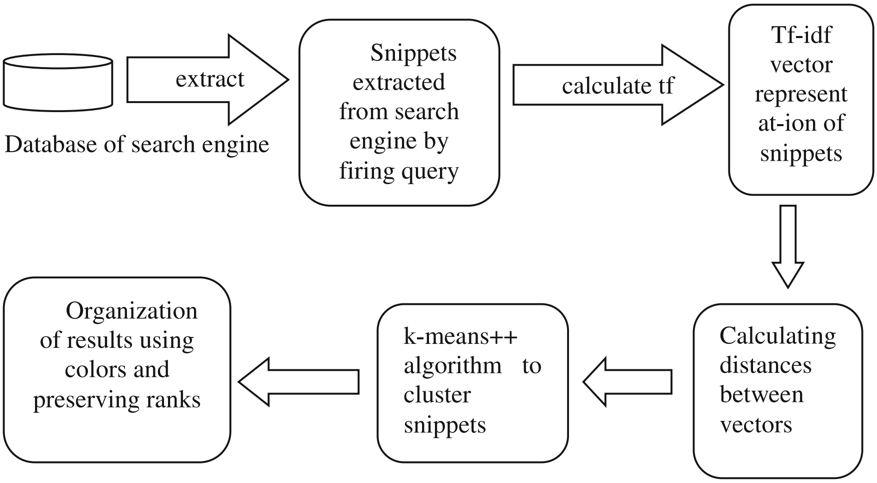
extract
Database of search engine
Snippets extracted from search engine by firing query
calculate tf
Tf-idf vector represent at-ion of snippets
Calculating distances between vectors
k-means++ algorithm to cluster snippets
Organization of results using colors and preserving ranks

Fig. 2 Architecture flow diagram

First, preprocessing of the snippets is done. The results are obtained through API of a web search engine by firing a query and passing it. The results are then extracted and stopword removal and stemming are done. Then, the tf-idf vector representation of each snippet is obtained.

Further, each vector of snippet is treated as a point and distances between each is calculated. Following this step, the vectors are then clustered by using a k-means ++ [9] algorithm that uses an automatically detects the centers unlike k-means.

Lastly, these clustered snippets are reprojected by giving colors to them according to the clusters to which they belong (Fig. 2).

5 Conclusion

We have proposed a technique that uses clustering on the snippets to group them according to a certain similarity based on the keywords of the query. Clustering and visualization on the snippets make it user friendly, easy, simple, and efficient. The use of k-means++ instead of k-means helps in improvising efficiency. The limitation is that it is useful for ambiguous queries rather than unambiguous ones.

References

1. Gomez-Nieto, E., San Roman, F., Pagliosa, P., Casaca, W., Helou, E.S., de Oliveira, M.C.F.: Similarity preserving snippet-based visualization of web search results. IEEE Trans. Vis. Comput. Graph. **20**, 457–463 (2014)
2. Heimonen, T., Jhaveri, N.: Visualizing query occurrence in search result lists. In: Proceedings of the Ninth International Conference on Information Visualisation, pp. 877–882 (2005)
3. Hearst, M.A.: TileBars: visualization of term distribution information in full text information access. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 59–66 (1995)
4. Kuo, B.Y.-L., Hentrich, T., Good, B.M., Wilkinson, M.D.: Tag clouds for summarizing web search results. In: Proceedings of the 16th International Conference on World Wide Web (WWW), pp. 1203–1204 (2007)
5. Hoeber, O., Yang, X.D.: The visual exploration of web search results using Hotmap. In: Proceedings of the 10th International Conference on Information Visualization, pp. 157–165 (2006)
6. Spoerri, A.: RankSpiral: toward enhancing search results visualization. In: Proceedings of the IEEE Symposium on Information Visualization, pp. 208–214 (2004)
7. Clarkson, E., Desai, K., Foley, J.: Resultmaps: visualization for search interfaces. IEEE Trans. Vis. Comput. Graph. **15**(6), 1057–1064 (2009)
8. Govardhan, P.H., Wagh, K.P., Chatur, P.N.: Web document clustering using proposed similarity measure. In: National Conference on Emerging Trends in Computer Technology, pp. 15–18 (2014)
9. Arthur, D., Vassilvitskii, S.: k-Means++: the advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete algorithms (SODA), pp. 1027–1035 (2007)

10. Nizamee, M., Shojib, M.: Visualizing the web search results with web search visualization using scatter plot. In: Proceedings of the IEEE Second Symposium on Web Society, pp. 5–10 (2010)
11. Sawaitul, S.D., Wagh, K.P., Chatur, P.N.: Classification and prediction of future weather by using back propagation algorithm-an approach. Int. J. Emerg. Technol. Adv. Eng. **2**, 110–113 (2012)
12. Dehankar, S., Wagh, K.P.: Web page classification using apriori algorithm and Naive Bayes classifier **3** (2015)
13. Charate, R.G., Chatur, P.N., Wagh, K.P.: Document filtering: intelligent inference system for web. Int. J. Manag. IT Eng. **3**, 207–222 (2013)
14. Kolhe, S., Wagh, K.P.: Semantic similarity based on information content. Int. J. Comput. Sci. Appl. 82–86 (2010)