A Review of Merging based on Suffix Tree Clustering

Priya S. Pujari
M.E student, Department of C.E
Dr. D.Y. Patil School of
Engineering & Technology
Savitribai Phule Pune University
Pune

Arti Waghmare
Assistant Professor
Dr. D.Y. Patil School of
Engineering & Technology
Savitribai Phule Pune University
Pune

ABSTRACT
Now a days web is growing very quickly, therefore it is important to search useful Patterns for web search document. The clustering of web search result has become a very interesting and popular research area to many organizations as it provides useful insights to information retrieval.Clustering of web search result system provides the search result for the user very concise and accurate, also provides reviews on them and locate specific information of interest.Clustering techniques can be used to organize retrieved results into a set of group based on their similarities. Several approaches existed for web document clustering using a suffix tree, but results show there is more research remains to be done.This paper presents a various approaches to Suffix Tree Cluster merge techniques to generate the informative, meaningful cluster and also compare some of the important cluster merging methods according to its performance .The method to merge clusters is used to sort out the problem of merging boundaries. Identical clusters are merged together based on a given estimation criterion until no more clusters can be merged. This survey aims to provide useful guidance for many applications where merging is having a remarkable impact on clustering.

General Terms
Web mining, Suffix tree, Clustering, Merging algorithm, Label

Keywords
Web, Information Retrieval, Grouping, Similarity, Search Result, Snippets

1. INTRODUCTION
In this paper, a brief survey of various suffix tree clusters merging techniques was introduced. As the web is growing very fast and web search engine contains a very huge amount of information, the results returned by the web search engine in response to the user query do not contain very useful information because of ambiguity in user query and it is also time consuming to search the informative documents.To overcome this problem, the approach of suffix tree clustering and merging is used. Suffix tree [1] is a data structure that keeps the set of text strings and containing all the suffixes for the given text also their values as position of text. Suffix tree efficiently determines documents which share commonphrases to form clusters.Any clustering technique relies on four concepts: data representation model, similarity measure, clustering model and clustering algorithm[2].The similar documents are grouped together.As the documents are being processed, the suffix tree is also expanded and after accessing each node, a list of documents is obtained along with the index. To increase the number of documents in each cluster, the similar cluster gets merged and merging is proceeding until no more clusters are found. The quality of a cluster depends on discovering hidden patterns.

For making search results very effective and informative, there are various merging techniques are available: Khoja stemmer for Arabic language browsing [3], Hownet [4] etc.

1.1 Motivation
Searching for information on the web is very time consuming process so to reduce the user's effort to gain the information quickly and to perform the information grouping manually, this survey present different approaches of merging in suffix tree clustering which gives very meaningful and informative clusters to satisfy the user query.

2. METHODS FOR MERGING

2.1 Cosine Similarity and Overlap Percentage
The paper [5] proposed a new technique for searching the knowledge about the topic or sub topics in deep, so that users can find their relevant topics systematically.First the technique searches the salient concept of topic from the documents given by the search engine and then find out informative pages which contain the topic with its salient concept but this approach fails to organize documents effectively and efficiently.To overcome this, cluster the documents properly. Sometimes all the documents in two clusters are similar but it is impossible to merge them because overlapping is occurring between them and it also produces too many clusters. To solve this problem, Jiangua Wang, Ruixu Li in 2006 introduced a novel cluster merging approach which will combine the cosine similarity and overlaps percentage and increase the efficiency of web search results [6].This method considers the similarity of non-overlap documents with adjusting to some parameters, namely k and α, also control the number of final clusters.These methods introduced the well known cosine similarity to the merging algorithm and evaluate the similarity between different clusters.

The cluster merging algorithm works as follows:

1. Search result obtained in response to user query was preprocessed, and then constructs the suffix tree, computing pairwise similarities among the cluster and taking the maximum.
2. After that, merge suffix tree clusters according to their similarity measure using a value of k then merge the base clusters with different value. This process continues until maximum less than the value of k
3. When merging completes, calculate the average number of final merged cluster.

This approach can also calculate the average number of merged clusters for different value of k. Here looked only at the cluster labels, ignoring the contents of each

cluster.Documents appearing in multiple clusters were removed from all but retained in highest ranked cluster and avoiding duplicates.Revise the overlap percentage calculation method to better reflect the overlap between two clusters.Then, employ the cosine similarity to calculate the similarity between the non-overlap parts.

2.2 Phrase Based Cluster Merging

A new similarity criterion for merging proposed by the author Kale A, Bharambe. U, M. SashiKumar [7] in 2009. In this approach, there are two important steps, namely Phrase Cluster Identification and Phrase Cluster Merging and described as follows:

1 Phrase Cluster Identification - Base clusters are identified for every branch of suffix tree and removing all the subsets from it, but retaining only one appearance of the phrase which is considered as a phrase label for that particular cluster.A follow-up work [8] showed how to avoid few suffix tree clustering limitations. Suffix tree recognizes inconsistent length phrases, but remove lengthy high quality phrases [9] and N-gram technique is used to find out fixed length phrases. By using both these methods, continuous and complete phrase label is identified. [10]

2 Phrase Cluster Merging - Base clusters are merged based on the overlap of their document sets as it reduces the number of clusters and avoid the overlap. If numbers of clusters are huge, then the ranking of the cluster is done using a score of base cluster and highest score displayed first in search results. For combining base cluster, Zamir. O and Etzioni. O introduced binary similarity measure criteria [11] but considered longest phrase only and not a short one. So to overcome this drawback new approach is introduced [2] which uses join operator to form a new common phrase of the cluster, when merging two similar base clusters is carried out.

These two steps are described in figure 1.This approach is also used in [12],[13]. It improves the quality of the cluster and use the frequency based approach to get descriptive phrases.

2.3 Suffix Tree Clustering With Label Merging

This approach is very useful in online social medium such as twitter which share and post news.The problem for social media is that it refers to the same topic many times and search result produces a long list of results.To overcome this problem new method called Suffix Tree Clustering and Label Merging is used [14].This approach merges partially overlapped labels

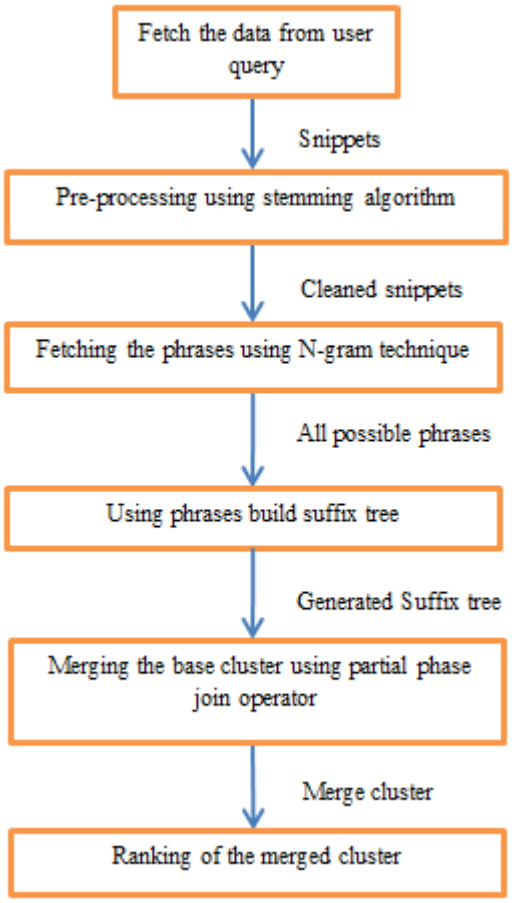

Figure 1: Phrase based cluster merging

and then combined into one label. (For example, the word fruit will be merged into mango) So by using this method, organize the tweets into a list of cluster labels for easy access of information to the user is possible.This approach generates a meaningful cluster label.First collect the tweets from many twitter profile pages and parsed, then convert given text into word token, removing the stop word, also check for duplicate tweets.After all this refinement of data, cluster labels are generated.

Then merging of label is working as follows and shown in following figure 2.

1 First taking the cluster label, then merge. The length method return number of words of every cluster label and create two-level cluster label structure.

2 Each cluster label is converted from shortened form full word form.

3 By using Part of Speech for every word in cluster label, removing the label which contains unmeanigful words and reduces the cluster size if cluster label is partially overlapped with another label, then it will be merged into one cluster and create two level cluster label structure.

This approach increased the performance using F1 measure.Still one issue, clustering of short text is not considered in this approach because tweets are also considered as short texts [15].Rangrej et al [16] studied various clustering techniques on a short text documents and provided experimental results using corpus from Twitter. Also merging based on improved hierarchical agglomerative

clustering is used for full subtopic retrieval, which retrieve more informative subtopics.It uses a cluster label based on key phrases with similar meaning and then merge these cluster label using this method [17]

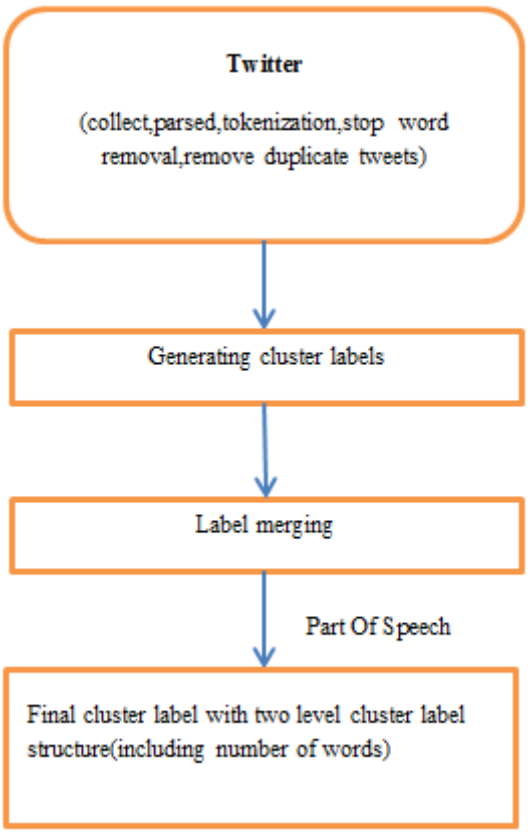

Figure 2: Cluster label merging

2.4 Merging the Semantic Duplicate Clusters

Suffix tree clustering (STC) is fast growing clustering algorithm [18] but sometimes problem occurs because of semantic duplicate cluster and some cluster contain the structure of another.The cluster displayed similar phrases which prevents other important phrases from being displayed [19].To identify these important phrases and to improve the quality of STC result, merge the semantic duplicate clusters also hierarchicalizing the label-contained clusters [20].This approach is very useful in dynamic clustering interface to web search results.It improves the organization of clustering results and better classification and Kohonen's feature semantic map is used for visualization of semantic relationships between input documents [21]

3. DISCUSSION AND CONCLUSION

In this paper, various merging techniques using a suffix tree clustering was studied.Merging is a very important step in information retrieval as it improves the quality of the cluster and ranking the cluster label results according to their frequency which removing the label which contain unmeanigful words.It also reducing the time complexity for searching the information so it is required to pay more attention towards improving merging process.Here this paper presents review on various merging techniques such as Overlap Criteria, Cosine Similarity, label merging etc. for getting more important clusters which will satisfied the user query quickly.

4. ACKNOWLEDGMENTS

The work described in this paper is supported by the Department of Computer Engineering of D.Y.Patil School of Engineering and Technology, Pune.The authors would like to thank to the principal Mr. Uttam B. Kalwane, and all staff members of Computer Engineering department for their valuable support.

5. REFERENCES

[1] E. Ukkonen, "On-line construction of suffix trees," Algorithmica, vol. 14, 1995, pp 249-260.

[2] S. Osi´nski, J. Stefanowski, and D. Weiss," Lingo: Search results clustering algorithm based on singular value decomposition," In Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference, Advances in Soft Computing, pages 359–368, Zakopane, Poland, 2004.

[3] S. Khoja , 1999. Stemming Arabic Text

[4] Zhao, P. And Cai, Q.S," Research Of Novel Chinese Text Clustering Algorithm Based On Hownet," In Computer Engineering And Applications, 43, 162-163,2007

[5] Liu B., Chin C. W., and Ng, H. T. Mining Topic-Specific Concepts and Definitions on tlie Web. In Proceedings of the Twelfth International World Wide Web Conference (WWW'OS), Budapest, Hungary. 2003

[6] Jianhua Wang, Ruixu Li," A New Cluster Merging Algorithm Of Suffix Tree Clustering," In International Federation For Information Processing, Volume 228, pp. 197-203, 2006

[7] Archana Kale, Ujwala Bharambe, M. Sashikumar," New Suffix Tree Similarity Measure And Labeling For Web Search Results Clustering," In 2nd International Conference On Emerging Trends In Engineering And Technology, pp 856 –861,Dec 2009

[8] Ferragina, P. and Gulli, A. (2004) "The Anatomy of a Hierarchical Clustering Engine for Webpage, News and Book Snippets," Technical report, RR04-04 Informatica, Pisa.

[9] Jongkol Janruang, Worapoj Kreesuradej, "A New Web Search Result Clustering based True Common Phrase Label Discovery," International Conference on Computational Intelligence for Modeling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), 2006.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn,"Data clustering: A review," ACM Computing Surveys (CSUR), 31(3),pp 264–323,1999

[11] Zamir, O. & Etzioni, O. (1998),"Web Document Clustering: A Feasibility Demonstration," Proc ACMSIGIR'98, Melbourne, Australia, 46-54.

[12] Aditi Chaturvedi, Dr. Kavita Burse, Rachna Mishra," Affinity Propagation Based Document Clustering Using Suffix Tree," In IJERT, Vol. 3, Issue 1, Jan 2014

[13] Hung Chim and Xiao tie Deng," Efficient Phrase-Based Document Similarity for Clustering," in IEEE Transaction on knowledge and data engineering , VOL. 20, NO. 9, Sept 2008

[14] Thaiprayoon S, Kongthon. A , Palingoon P And Haruechaiyasak.C," Search Result Clustering For Thai Twitter Based On Suffix Tree Clustering," In 9th International Conference , pp 1 – 4, May 2012

[15] F. Perez-Tellez and D. Pinto, "Om the Difficult of Clustering Company Tweets," Proc. Of the 2nd International Workshop on Search and Mining User-generated Content (SMUC '10), pp. 95–102, 2010.

[16] A. Rangrej, S. Kulkarni and A. V. Tendulkar, "Comparative Study of Clustering Techniques for Short Text Documents," Proc. Of the 20th International World Wide Web Conference (WWW, '11), pp. 111–112, 2011

[17] Andrea Bernardini, Claudio Carpineto, Massimiliano D'amico," Full-Subtopic Retrieval With Keyphrase-Based Search Results Clustering," In IEEE/WIC/ACM International Conferences On Web Intelligence And Intelligent Agent Technologies-Workshops,Volume 1,Pp 206 – 213,Sept 2009

[18] R.Mahalakshmi, V.Lakshmi Praba," A Relative Study On Search Results Clustering Algorithms - K-Means, Suffix Tree And Lingo," In International Journal Of Engineering And Advanced Technology, Volume-2, Aug 2013.

[19] Oren Zamir ,Oren EtzioniO. "Grouper: A Dynamic Clustering Interface to Web Search Results," University of Washington. Department of Computer Science and Engineering. 1999K. Elissa

[20] Guodong Hu, Wanli Zuo, Fengling He,Ying Wang," Semantic-Based Hierarchicalize The Result Of Suffix Tree Clustering," Proc In Second International Symposium On Knowledge Acquisition And Modeling, Volume 03,pp 221-224 , 2009

[21] Lin, X., "A self-organizing semantic map for informationretrieval," Proceedings of the 14th International ACM SIGIRConference on Research and Development in Information Retrieval (SIGIR'91), 1991, pp. 262-269.