ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 7, July 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Clustering uses Modified Approach for Content and Side Information

Snehal S. Ingavale[1]
Computer Department
TSSMs Bhivarabai Sawant College of Engineering and Research
Pune, India

Prof. M. K. Kodmelwar[2]
Computer Department
TSSMs Bhivarabai Sawant College of Engineering and Research
Pune, India

Abstract: ***In various applications of text mining, each text documents having side information. The text document contains side information such as the links in the document, user-access behavior from web logs, document provenance information or other non-textual attributes. This attributes may contain a lot of information which is used for clustering purposes. In any case, with respect to significance of this side-information may be hard to assess, particularly when a percentage of the information is noisy. Existing framework proposes established probabilistic models with partitioning algorithms keeping in mind the end goal to make a successful clustering approach. In proposed system the term frequency cosine angle based similarity is measuring for Content as well as for side information. Implements a modified approach for COATES using cosine angle based similarity measure.***

Keywords: Text Mining, Classification, Clustering, Side Information, COATES.

I. INTRODUCTION

Data Mining is known as the type of database analysis that attempts to extract useful patterns or relationships in a group of data. A major goal of data mining is to extract previously unknown useful relationships among different data.

A. Text Mining

The text clustering issues raises in the context of numerous application areas for example social networks, the web and computerized accumulations. Text information is the quickly expanding measures of the vast online accumulations have prompted an enthusiasm for making adaptable and powerful algorithms which are used in mining process. In recent years a large amount of work has been done on the problem of clustering in text document collections in the database and information retrieval. However, the problem of pure text clustering in the absence of other kinds of attributes this work is primarily designed. In many application domains, a tremendous amount of side information is available with each document. This is because text documents having the different types of applications in which there may be a large amount of other kinds of database attributes or Meta information which may be useful into the clustering process [1].

B. Side Information

The problem of text mining arises in the context of many application domains such as the web, social networks, and other digital collections. Some examples of such side information are as follows:

1. Web logs

In an application in which we track user access behavior of web documents, the user-access behavior may be captured in the form of web logs. For each document, the meta-information may correspond to the browsing behavior of the different users. Such logs can be used to enhance the quality of the mining process in a way which is more meaningful to the user, and also

application-sensitive. This is because the logs can often pick up subtle correlations in content, which cannot be picked up by the raw text alone.

2. Links present in Text Document

Text documents, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes. As in the previous case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.

3. Meta-data

Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the document. In other cases, data such as ownership, location, or even temporal information may be informative for mining purposes. In a number of network and user sharing applications, documents may be associated with user-tags, which may also be quite informative.

II. RELATED WORK

Charu C. Aggarwal, Yuchen Zhao and Philip S. Yu [1] designed an algorithm which combines classical Partitioning algorithms with probabilistic models in order to create an effective clustering approach. They presented experimental results on a number of real data sets in order to illustrate the advantages of using such an approach. They presented methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side-information or Meta information, which might be used in order to improve the clustering process.

S. Guha, R. Rastogi, and K. Shim demonstrates [2] that for discovering groups and identifying interesting distributions in the underlying data clustering is used in data mining. Traditional clustering algorithms either favor clusters with spherical shapes and similar sizes. In this paper a clustering algorithm is presented which is called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size.

D.Cutting, D. Karger, J. Pedersen, and J. Tukey [3] explains the Hybrid Technique (Scatter-gather technique is the hybrid clustering technique). An example of the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered document collection of the document organization. Initially the system scatters the collection of document into a small number of several document groups, or clusters, and presents short summaries of documents to the users.

T. Liu, S. Liu, Z. Chen, and W.Y. Ma[4] explains Feature extraction and feature selection techniques are used to reduce feature space dimensionality. In feature extraction it extracts a set of new features from original features through some functional mapping. In feature selection it chooses a subset from the original feature set according to some criteria. Document frequency, information gain, term strength are some of the feature selection methods.

S. Zhong demonstrates [5] that clustering data streams has been a new research topic, recently used in many real data mining applications, and has attracted a lot of research attention. However, there is not much work on clustering high-dimensional streaming text data. This paper merges an efficient online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams.

III. EXISTING SYSTEM

In existing system Cosine similarity is used for document content clustering and Posterior probability is used for side information clustering. In the existing system, side-information combines into the mining process is very dangerous, because it can either add noise to the process or it can improve the quality of the representation for the mining process.

IV. PROPOSE SYSTEM

In propose system a modified COATES algorithm is used for efficient clustering approach. The term frequency cosine angel based similarity is calculated for content as well as for side information, to improve the clustering process.

As show in Figure 1 the system architecture consist of modules Preprocess dataset, Select attributes and mining and Clustering. In Pre-process dataset module take the input dataset and remove null and abnormal data from it. Re-move the null value from the dataset which is not useful and the meaningless for the auxillary attributes. In Select Attribute and mining module apply the Stopword removing algorithm for remove the stopwords in the document. Stopword means a, an, the, are, of etc. Then apply the Stemming algorithm for convert the variant form of word into the normal form. Stemming means remove the ed, tion, es etc. Attribute selection is important process in the system. Appropriate attribute must be selected.

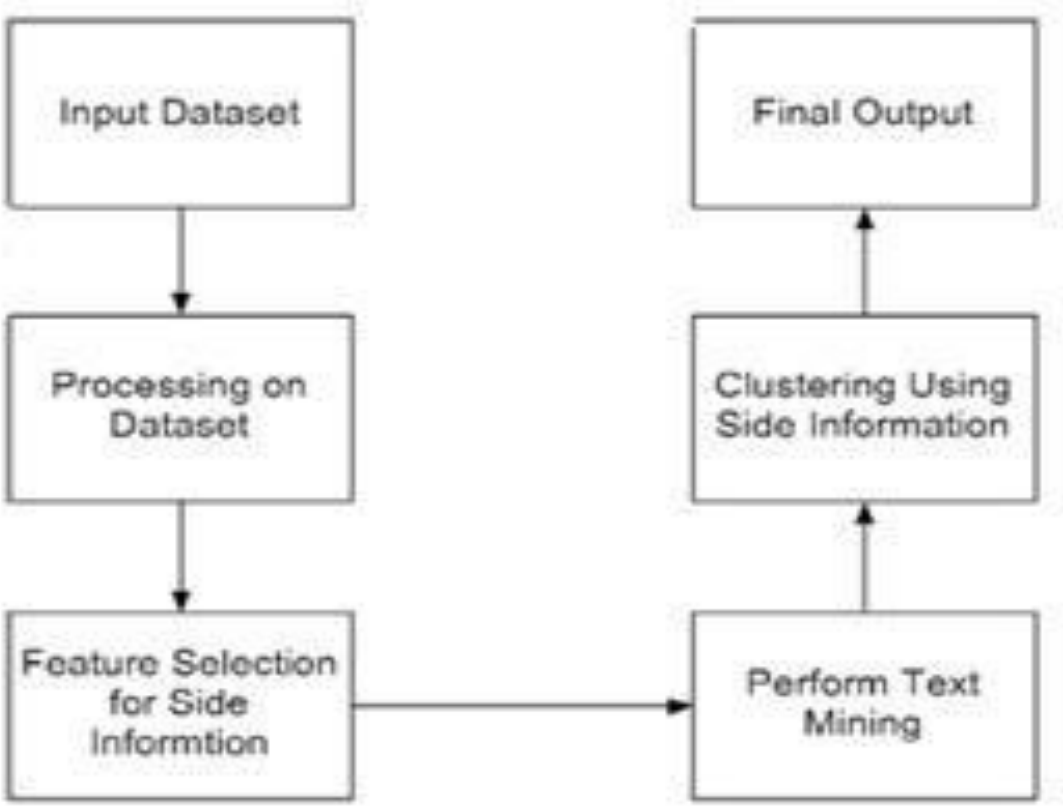


Fig.1 System Architecture

A. *Similarity based on Term Frequency*

Tf-idf, is the short form of the term frequency inverse document frequency, is a numerical calculation for each word in the document to show that how important a in a collection or corpus. Weighting factor is used in information retrieval and text mining. The number of times a word appears in the document increases then proportionally increase the Tf-idf value, yet is offset by the frequency of the word in the corpus, which adjusts for the fact that a few words appear all the more as often as possible in general. One of the most straightforward summing so as to rank functions is processed the tfidf for each term; many more sophis-ticated ranking functions are variants of this basic model.

B. Flowchart

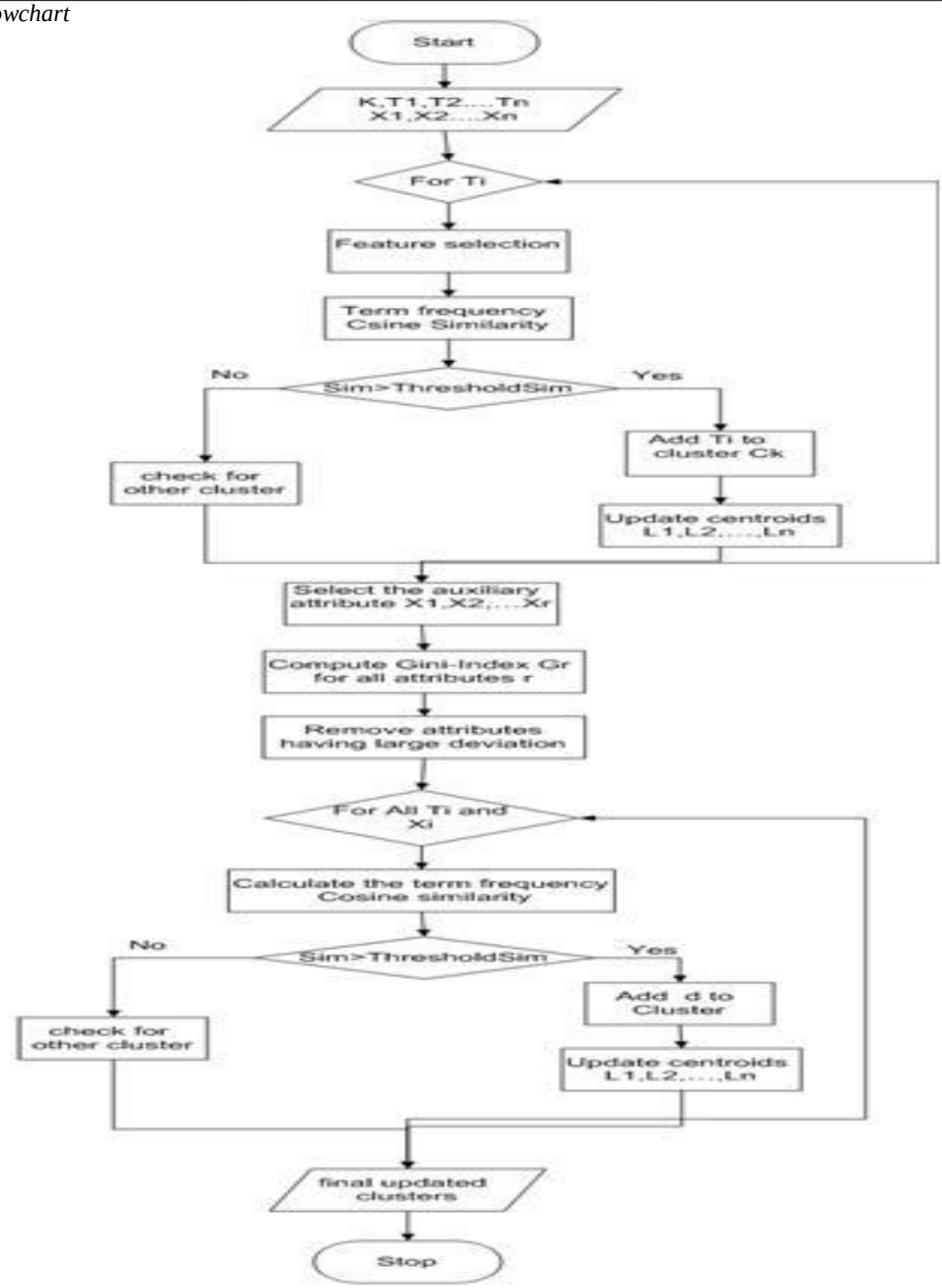


Fig.2. Flowchart

Figure 2 shows the proposed systems flow chart. This flow chart contains Add cluster, Update Centroid, compute Gini index, Term frequency Cosine Similarity functions. These functions are used for clustering the document.

C. Gini Index

Gini index need to dispose of the noisy at-tributes. This is particularly vital, when the quantity of auxiliary attributes is very huge. In this manner, the start of every auxiliary iteration. Computation of the gini-index is depends on the clusters made by the last content based iteration. This gini-index gives an evaluation of the biased force of every credit as for the clustering process.

V. MATHEMATICAL MODEL

Mathematical model of a proposed system is given below.

Let S be the proposed system, I is an input set to the proposed system, and O represents the output of the system.

Where

S = {S0; Sf; I; O; F; SF; DD; NDD}

S0 = {Initial state}

Sf = {Final state of clustered data}

I = {Data set D}

D = {x1,x2,xn}

Xi = {Data item x}

O = {Cn number clusters of D}

Ci = {Cluster c}

n = {Number of clusters}

F = {GetDataSet, GetClusters}

SF = {GetSimilarity,GetGiniIndex}

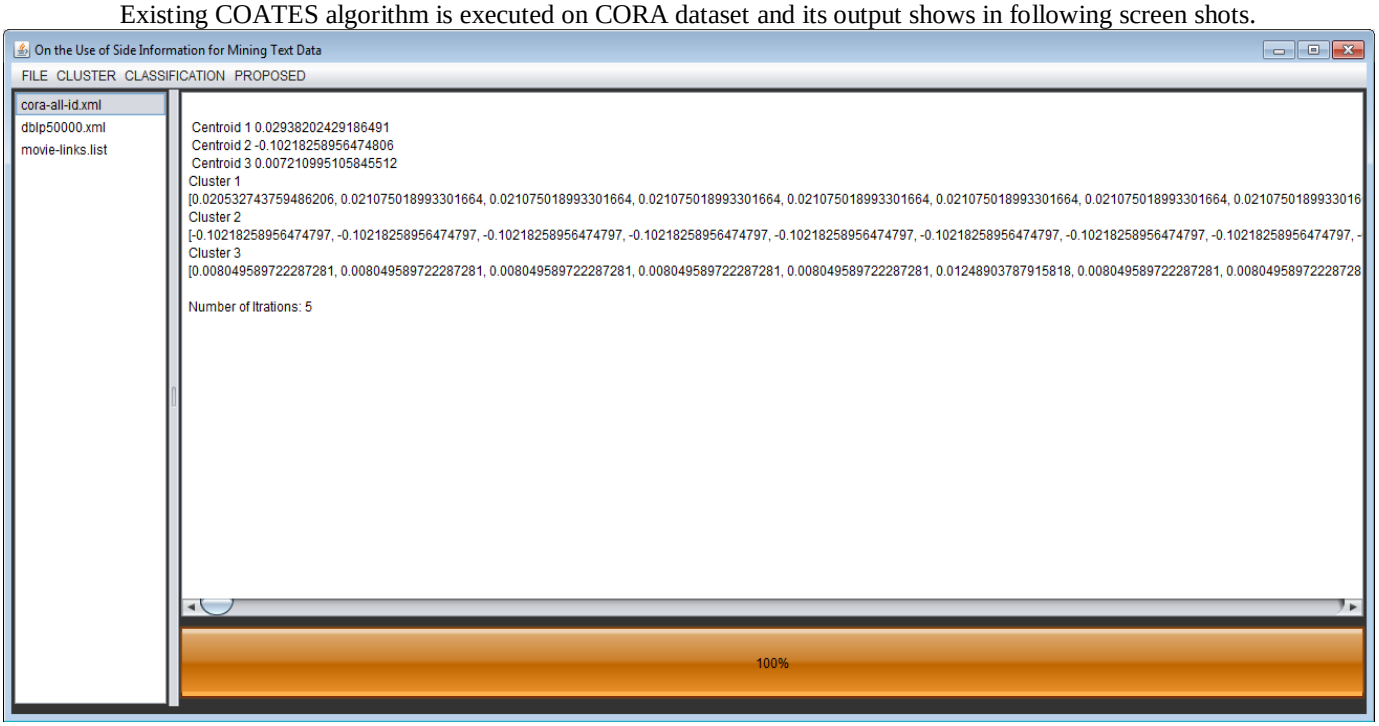
DD = {Clustered Data D}

NDD = {Non deterministic data}

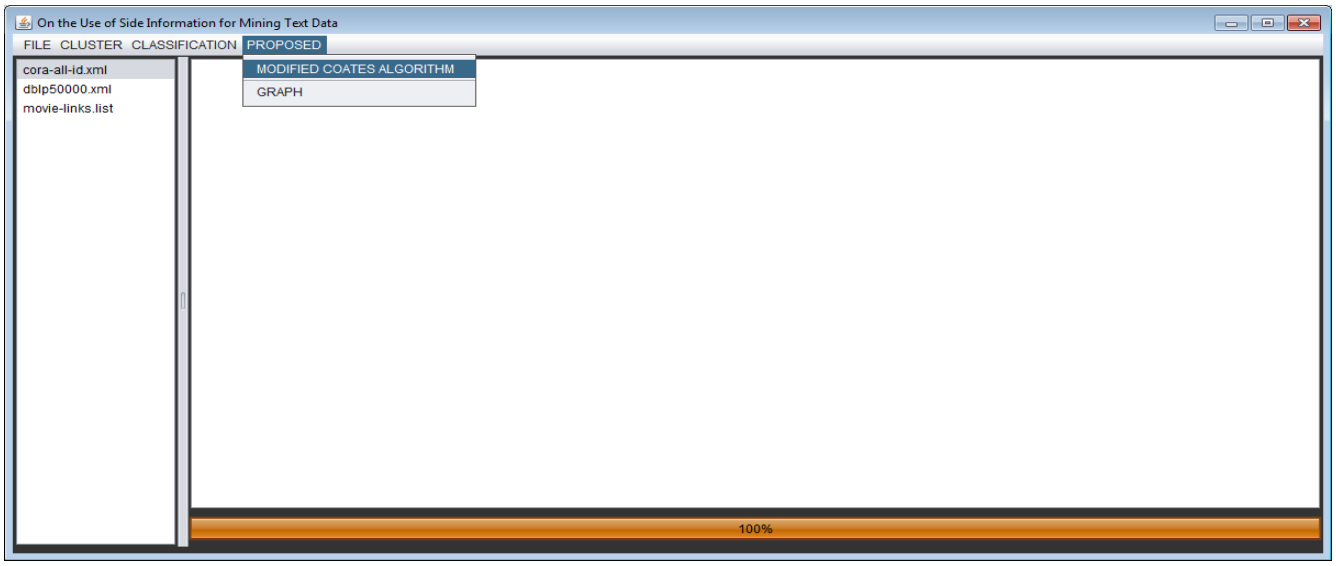
VI. EXPERIMENTAL RESULT

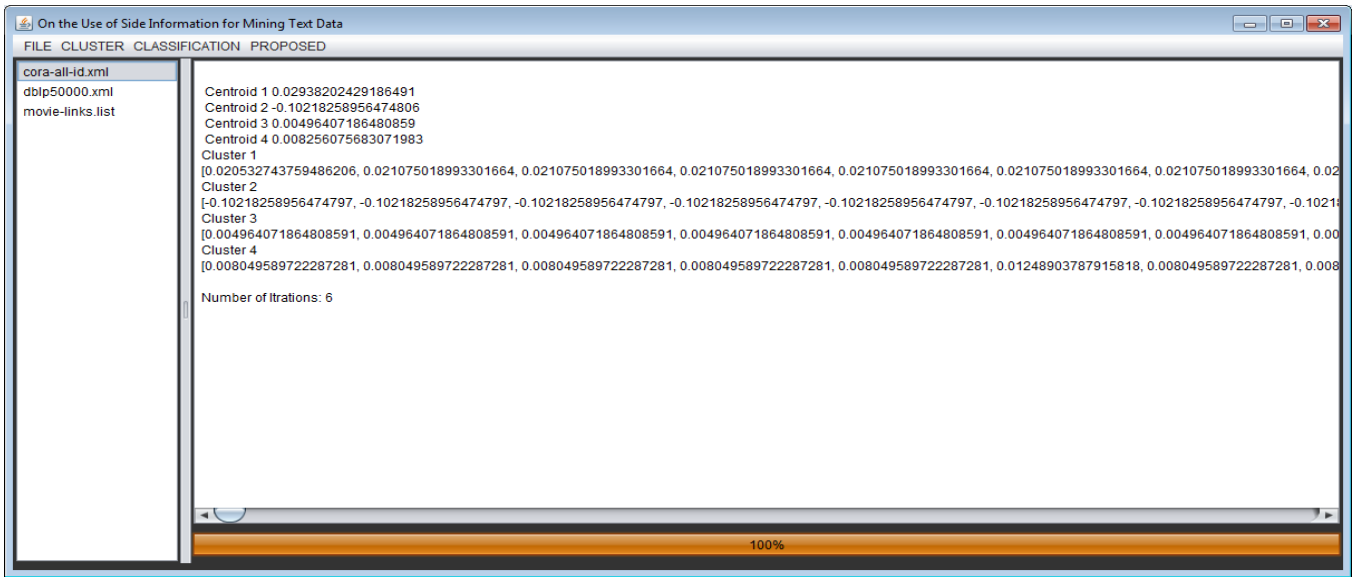
Propose framework use the text preprocessing methods, for example, stops words removing, stemming, Term Frequency computation and the time required for side-information preprocessing is insignificant. These running times just for the clustering portions keeping in mind the end goal is to sharpen the comparisons and make them more meaningful. The modified COATES algorithm is used to combine the text and side information into the clustering process, it is very important way than others, its running times are anticipated that would be more. The goal of modified COATES algorithm is, better qualitative results with less running time. Also tried the effectiveness of the strategy with increasing data size. This is finished by sampling a portion of the data, and reporting the results for various sample sizes.

Existing COATES algorithm is executed on CORA dataset and its output shows in following screen shots.

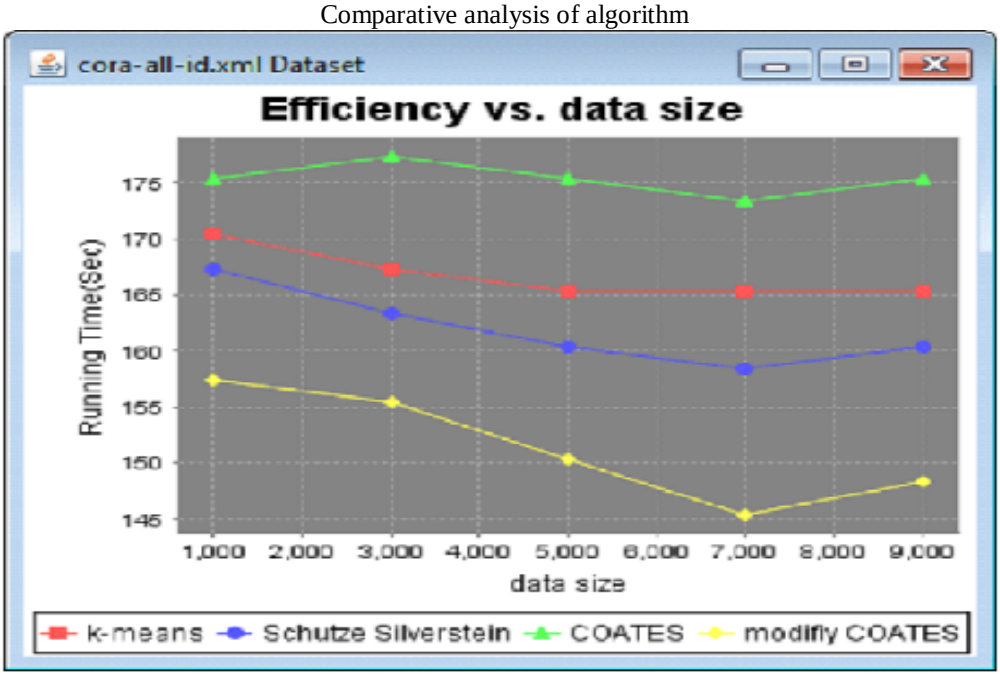


Modified COATES algorithm is executed on CORA dataset and its output shows in following screen shots.





Comparative analysis of algorithm



VII. Conclusion

In this paper, side-information is used with text data and performs mining process. A large amount of side-information or meta-information is available with different text documents or database. This side information is used to improve the clustering process. The term frequency cosine angel based similarity is calculated for content as well as for side information to improve the clustering process. By using side-information can easily increase the quality and efficiency of text clustering.

Acknowledgment

I have taken efforts in this review of clustering for mining using side information. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Prof. D. O. Shamkuwar for his guidance and constant supervision as well as for providing necessary information regarding this approach.

References

1. Aggarwal, C.C., Yuchen Zhao, Yu, P.S."On the Use of Side Information for Mining Text Data", Knowledge and Data Engineering, IEEE Transactions on, Vol. 26, No. 6, pp. 1415-1429, June 2014.
2. S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large Databases,” in Proc. ACM SIGMOD Conf., New York, NY, USA, 1998, pp. 73–84.
3. D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.
4. T. Liu, S. Liu, Z. Chen and W.Y. Ma, “An evaluation of feature selection for text clustering, In Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488 – 495.
5. S. Zhong, “Efficient streaming text clustering,” Neural Netw., vol. 18, no. 5–6, pp. 790–798, 2005.
6. S. Guha, R. Rastogi, and K. Shim, “ROCK: A robust clustering algorithm for categorical attributes,” Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.
7. G. Karypis and Y. Zhao,”Topic-Driven Clustering for Document Datasets”, Proc. SIAM Int’l Conf. Data Mining, pp. 358-369, 2005.
8. Frigui H and Nasraoui O.,”Simultaneous clustering and dynamic keyword weighting for text documents” M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45-70.
9. F. Sebastiani, ”Machine learning for automated text cate-gorization” ,ACM CSUR, vol. 34, no. 1, pp. 1-47, 2002
10. Zhong S, ”Efficient streaming text clustering” ,Neural Netw., vol. 18, no. 5-6, pp. 790-798, 2005.

Author(s) Profile

Snehal Ingavale, Is currently pursuing M.E (Computer) from Department of Computer Engineering, TSSMs Bhivarabai Sawant College of Engineering and Research, Pune, India. Savitribai Phule Pune University. She received her B.E (Computer) Degree from Dnyanganga College of Engineering and Research, Pune, India Savitribai Phule Pune University. Her area of interest is data mining, web mining.

Manohar K. Kodmelwar, BE(CSE) ,ME Comp, Ph.D. (Pursuing). He is currently working as Asst. Professor with Department of Computer Engineering, Bhivarabai Sawant College of Engineering and Research, Pune, MAH, India.

