Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 807 – 814

Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

Text Mining using Metadata for Generation of Side information

Shraddha S.Bhanuse[a] ,Shailesh D.Kamble[b],Sandeep M. Kakde[c]

[a,b]*Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Hingna Road, Nagpur 441110*

[c]*Department of Electronics Engieering, Yeshwantrao Chavan College of Engineering, Hingna Road, Nagpur 441110*

Abstract

Text Mining is knowledge discovery process from large database to find out unknown patterns. In many metadata based text mining applications, side information also known as metadata which is associated with the text document. There are different types side information containing large amount of data i.e. metadata, weblogs and non-textual data (image, video, etc.).The side information is difficult to estimate when it contains noisy data. To achieve this, there is scope of improvement in generating side information i.e. selecting efficient classification and clustering algorithms, providing security for clustered side information, document organization, exploring filtering approaches. In future, there is a scope to design an extended approach for clustering using classical partitioning and probabilistic model.

Peer-review under responsibility of organizing committee of the ICISP2015

*Keywords:*Text Mining,metadata,side information,classification,clustering,filtering,security;

1. Introduction

In text mining applications [4], are the challenges of clustering text data remains in the text mining area such as internet technology, i.e .www, online resources dataset and the social networking sites. The issue of Text clustering in data processing arises in the surround of many application domains such as the web, Social network and online

* Corresponding author. Tel.: +91-8378968842
*E-mail address:*shradha.bhanuse@gmail.com

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.061

Digital collection. The rapid increasing amounts of text data in the surround of these large online collections from web, Facebook, LinkedIn and Twitter In the recent year, a lot of research work is carried out on clustering large text data in text mining area[15,16,17]. Therefore, from the studied existing literature, there is a major challenge in front of the research communities on creating scalable and effective mining algorithms. The examples of metadata based such side information are:

Metadata information available in the web logs and which gives information related to browsing behavior of various users. For improving the quality of the text mining we can track such web logs.

- Many text documents having connections among them are also called as attributes. Such attributes, link contains useful information for text mining purpose.
- Meta information associated with the text document consisting of different attributes i.e location, ownership, user tags in a network and other information related to the given original document may be important for text mining process. Metadata information also useful in online shopping and in business Intelligence.

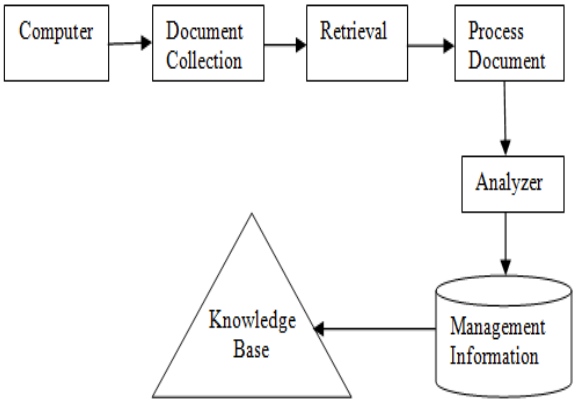

Fig.1. Flow Diagram of Text Mining

Text mining is new area of Computer science and Engineering and which can be support to natural language processing. Text mining extracts relevant information from unknown pattern [1]. Data processing can help to organize valuable business data from text-based content such as blogs, email, posting, and social media. Intelligent Text Analysis is also called as Text mining. Extracting interesting or uninteresting useful metadata based text information. The difference between text mining and data mining is that, data mining are used to process the structured data and metadata based text mining used to process the unstructured data [19].

For improving the quality of text data, metadata based text mining technique [2] is used. Therefore, it helps to get better accuracy and efficiency for clustering text data. In data processing, generating the feature vector is a special form of dimensional reduction. To transform data as input data to the set of features is called feature vector. This paper organizes are as follows: Section 1 gives overview about introduction of text mining. Section 2 gives the overview of side information .Section 3 describes the various classification and clustering algorithms for generating the side information and existing literature approaches. Section 4 describes the standard dataset overview and initial preprocessing in terms of different types of analysis. Discussion on metadata based text mining and further implementation process is described in section 5.

2. Side Information

Clustering with this noisy data is a challenge of data processing [28].Web logs: In many applications in which we track user access behaviour of web documents, the user-access behaviour captured in the form of web logs. Such logs can be used to improve the quality of data processing system[5]. Links present in Text Document: Text documents are also called as attributes [27, 17]. Such links contain a lot of useful information for data processing. Such attributes may often provide insights about the relation and correlations among documents in a way which may not

be easily accessible from raw content data [28, 23, 21]. Meta-data: Many web documents have meta-data associated with them .The meta- data correspond to different kinds of attributes such as the provenance or related or other information about the origin of the document[26], Example data such as ownership, location and user tags, etc

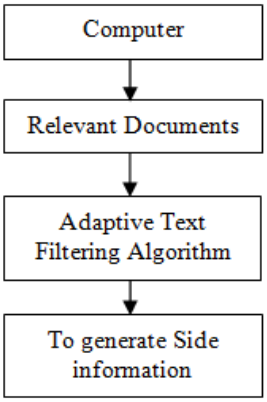

Fig.2. Generation of Side information

3. Related work

In metadata based text mining, large online collection is main reason to develop a mechanism to create effective and scalable clustering algorithms used for generating side information [15, 16]. The current approaches focuses on data processing to maximize the clustering advantage to generate side information. Guha S. et al. [6, 7] proposed an approach for clustering text data with side information. It gives a way perform the mining process as to maximize the benefits of [1] side information. It accomplishes in an algorithm which combines traditional partitioning algorithms with probabilistic models for effective clustering. Douglass R. et al. [8] suggested that for metadata based retrieval, document clustering has not been well used. Two main categories for its objection: first, for large clustering is too slow and second, that retrieval is not improved by classification and clustering [32, 33]. The clustering is used to improve predictive search techniques. However, data processing as an information right of entry tool in its own right avoid these objections, and provides efficient new access paradigm. Document clustering is presented as initial document browsing technique [29]. There is strong association rule between clustering and its technique. Meta information technique is used for feature compression and extraction of reduce dimensionality. Clustering is main challenge of metadata based text mining. We have different types of algorithm for classification and clustering. In text mining, data processing plays very important role. Aggarwal C.C. et al. [9] presented a survey on text data classification and clustering algorithm. For classification and clustering data is extracted or used from metadata for generating side information. Guha S. et al. [6, 7] presented the unknown discovery pattern or identifying interesting pattern in terms of data clustering is used in data mining. Clustering algorithm is implemented called as CURE which is more accurate to outliers, and identifies clusters. Because cluster having non-spherical shapes and wide variances in size. Zhong N. et al. [10] proposed an effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving. To enhance effectiveness of modifying the discovered patterns to find appropriate unknown pattern. Franz M.et al.[11] proposed unsupervised and supervised learning, will helps to improving the quality of the clustering effects of both the text metadata and side information. The proposed approach shows extension of the clustering approach to the metadata based text classification using the side information or generating side information of the text documents.

Clustering Algorithms

Hierarchical clustering algorithms form a cluster hierarchy or tree of cluster which contains parent node and child node. Sibling clusters partition their parent cluster property. K- medoid clustering algorithm is based on the iterative approach. The set of k representative are improved by using randomized method. Online spherical k-means

algorithm [24, 29] works with cosine similarity. It is well known approach for clustering high-dimensional data.

Classification Algorithms

Decision trees are used for the classification from hierarchical data Pattern classifier**;** we can analyze the word patterns, frequency count for metadata processing with different classes. For classification rules are used. SVM Classifiers are used for partition the data space[30]. Neural Network Classifier [23] is used for text classification[16]. Bayesian classifier [30] is based on the probabilistic approach to classify the text of word count in document. Naive Bays classifier [31] is a simple probabilistic classifier based Byes' theorem with strong (naive) assumptions.

3.1 Clustering Side information

The approach shows the advantages of using side-information for data processing [28, 24]. Content and auxiliary attribute based text Clustering (COATES) algorithm is used without any side information for performing initial text mining. Basically data processing work on Content and Auxiliary attributes. In data processing, COATES algorithm works in two steps: *Initialization:* It is first step in data processing .In this step basic clustering [27, 31]approach is used without any side information. In this step k-means algorithm is used for clustering. K-means is a simple algorithm which will work very quickly and efficiently. It provides start point. Partitioning provide only on the basis on information not on side information [26, 27, 14]. *Main Step:* This steps starts with partitioning approach and then grouping the most relevant documents. It is an iterative process. In data processing, text content and the auxiliary information helps to improve the quality of metadata based text clustering. Content based attributes presents exact or relevant information and auxiliary is not exact but related data.

3.2 Classification with side information

Clustering approach is used in classification of text based metadata using side information. Naive Bays Classifiers, Decision Tree Classifier, SVM classifiers are used for classification [23, 30]. By analysis on data processing, it creates a model which provides class distribution on basis of Training set model. It uses supervised clustering approach using side information. The Classification algorithm works in 3 steps. *Feature Selection:* Feature selection is used to remove to unwanted data. It works for both the content text attributes and the auxiliary attribute. *Initialization:* In initialization process, a supervised k means approach is used for separation of Content based and auxiliary text content. *Cluster-Training Model Design:* This step, combines the text and side information used in data processing for the creating a new cluster.

4. Datasets

CORA dataset: The Cora dataset is collection of set of rules and their relations. It allows performing various machine learning approaches. It is collection of number of authors and scientific papers. For Cora assigns to each paper a set of categories which are selected from taxonomy of classes and their relationship. The Cora dataset contains the publications data in the field of computer science and engineering. In this paper the database of each paper is classified by topics i.e. Artificial Intelligence, Data Structures Algorithms and Theory, Networking, Hardware and Architecture, Programming, Human Computer Interaction, Information Retrieval, Operating Systems, Databases and Encryption and Compression. Word dictionary is used, which is related to dataset. We are using SQL database for storing and retrieving information.

IMDB and DBLP data set: IMDB is an internet movie database is collection of online digital information. Dataset contains 10years data. It contains class labels Drama, Comedy, Short, documentary. DBLP-Four dataset .We performs clustering on the combinations, actor, co-actor, director, producer and obtained result on the basis of accuracy, no. of cluster.

In data processing, analysis is the process that will convert a stream of characters into streams of words [32]. First it will perform tokenization and then stemming is performed where most redundant word are removed. Means on, is, was, were, in, out, above, etc. The words are removed from the documents .Experimentation are carried out on CORA dataset and analysis of text data carried out by using different parameters as shown in figure number 3.

- Count the frequency of words.

- Keyword Analysis
- Count the no. offset
- Count the no. of white space

Experimentation carried out on CORA dataset.

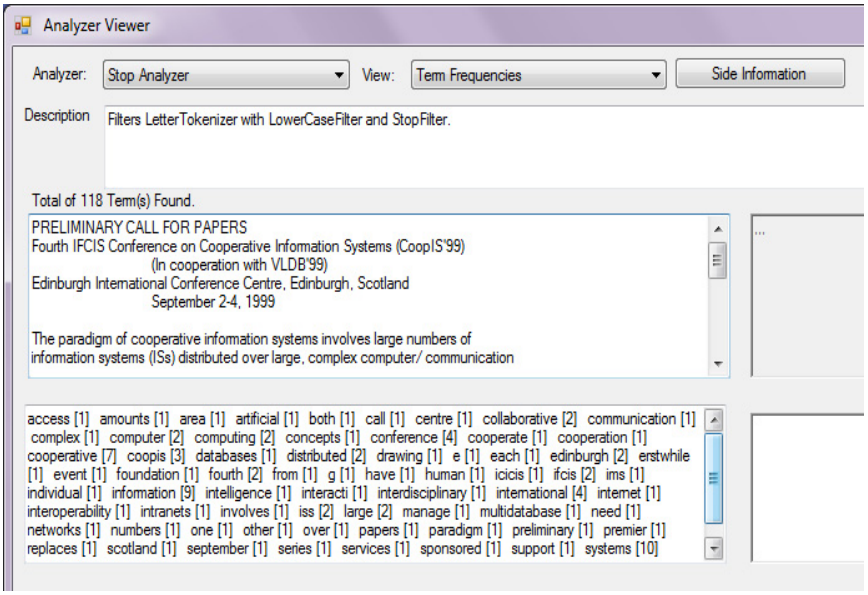

Fig.3. Analyzer for word count

Table1: Analysis of word count

Words	Frequency	Word	Frequency	Word	Frequency
Access	1	Area	1	database	1
cooperate	1	cooperative	7	internet	1
foundation	1	information	9	international	4
computing	2	conference	4	system	10

In this above table, proposed work helps to analyze the frequency of different words. Word frequency of access is 1.Word frequency of area is 1.Word frequency of cooperative is 7.Word frequency count of information is9.Word frequency of conference is 4.The frequency count of system is 10.This is helpful for generation of side information in the form of word indexing. Word indexing is useful for priority of words according set of rules. Example is Google for page ranking word index terminology is used.

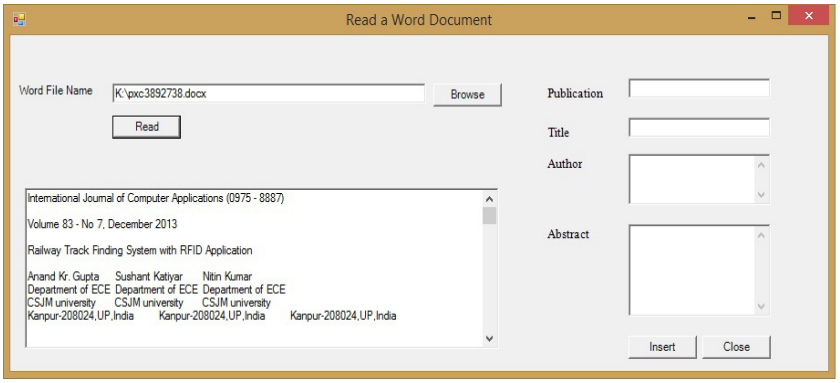

Fig.4. Generation of side information

In the above diagram, as per input is provided which is in the text file (conference paper) and output is generation of side information. Side information will be paper's publication, name, title, author, important concepts or terms from abstract and keywords. So generating side information is helpful in text mining using metadata.

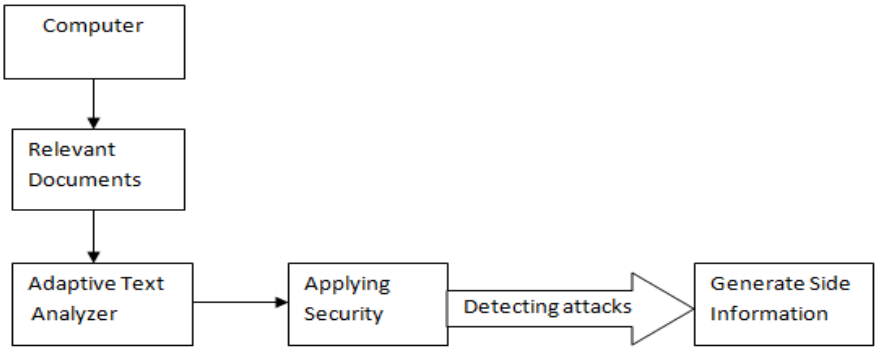

Fig.5. Providing security to side information by Intrusion Detection System

In above diagram, proposed approach collected relevant documents from computer. Relevant documents analyzes by Text adaptive algorithm. Intrusion Detection System is used for providing security for generating side information. DES is a block cipher encryption algorithm: it takes a fixed-length block of data and converts it into a fixed-length block of encrypted data of the same size by using a symmetric key. The key's length is 64 bits, but because 8 bits are used for parity, the effective key length is 56 bits. Decryption uses a reverse process on the encrypted data block with the same symmetric key, resulting in the original clear-text block of data.3DES uses three stages of DES and is more secure. DES is applied three times with three different 56-bit keys, resulting in an effective key length of 168 bits. Whereas no successful attack has ever been documented in cracking 3DES, this enhanced security of DES is sufficient for most current applications. So Intrusion Detection System detects attack and by analyzing the behavior of attack some countermeasure will apply. Secured side information is generated. This side information is useful for Text Mining.

5. Discussion

This paper gives idea about the use of metadata in text mining for generation of side information. Classification and clustering can be performed on the basis of side information different classification and clustering algorithms. Text mining uses metadata information for mining text data. To design clustering, classical partitioning uses probabilistic model to create effective clustering. Existing experimental results are measured in terms of number of cluster, running time and accuracy.

In future, there is a scope to design an extended approach for clustering using classical partitioning and probabilistic model. Also there is a scope in providing security for clustered side information and exploring the filter approaches implementation will work on training set model.

References

1. C. C. Aggarwal and C.-X. Zhai, *Mining Text Data*. New York, NY, USA: Springer, 2012.
2. C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*.New York, NY, USA: Springer, 2010.
3. 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.
4. C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams," in *Proc. SIAM Conf. Data.*
5. Mining, 2006, pp. 477–481.28 T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in Proc.
6. Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, Fello " On the Use of Side Information for Mining Text Data" IEEE Transactionson knowledge and data engineering vol 26,no.6 pp 1415-1429,2014
7. 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.
8. 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.
9. 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.
10. C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in *Mining Text Data.* New York, NY, USA: Springer,2012.
11. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu," Effective Pattern Discovery for Text Mining" IEEE *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.*
12. M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, "Unsupervised and supervised clustering for topic tracking," in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 2001, pp. 310–317.
13. H. Schutze and C. Silverstein, "Projections for efficient document clustering," in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1997,pp. 74–81.
14. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. Text Mining Workshop KDD*,2000, pp. 109–110.
15. S. Zhong, "Efficient streaming text clustering," *Neural Netw.*, vol. 18, no. 5–6, pp. 790–798, 2005.
16. R. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," in *Proc. VLDB Conf.*, San Francisco, CA,USA, 1994, pp. 144–155.
17. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM SIGMOD Conf.*, New York, NY, USA, 1996, pp. 103–114.
18. Y. Sun, J. Han, J. Gao, and Y. Yu, "TopicModel: Information network integrated topic modeling," in *Proc. ICDM Conf.*, Miami, FL,USA, 2009, pp. 493–502.
19. T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proc.ACM KDD Conf.*, New York, NY, USA, 2009, pp. 927–936.
20. Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *PVLDB*, vol. 2, no. 1, pp. 718–729,2009.
21. C. C. Aggarwal and P. S. Yu, "On text clustering with side information," in *Proc. IEEE ICDE Conf.*, Washington, DC, USA,2012.
22. R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in *Proc. CIKMConf.*, New York, NY, USA, 2006, pp. 778–779.

23. Dhillon, S. Mallela, and D. Modha, “Information-theoretic coclustering,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2003,pp. 89–98.
24. C. C. Aggarwal, *Social Network Data Analytics*. New York, NY, USA: Springer, 2011.
25. McCallum. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering* [Online].
26. G. Salton, *An Introduction to Modern Information Retrieval.* London,U.K.: McGraw Hill, 1983.
27. C. Silverstein and J. Pedersen, “Almost-constant time clustering of arbitrary corpus sets,” in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1997, pp. 60–66
28. W. Xu, X. Liu, and Y. Gong, “Document clustering based on nonnegative matrix factorization,” in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 2003, pp. 267–273.
29. Y. Zhao and G. Karypis, “Topic-driven clustering for document datasets,” in *Proc. SIAM Conf. Data Mining*, 2005, pp. 358–369.
30. C. C. Aggarwal, S. C. Gates, and P. S. Yu, “On using partial supervision for text categorization,” *IEEE Trans. Knowl. Data Eng.*,vol. 16, no. 2, pp. 245–255, Feb. 2004.
31. A Jain and R. Dubes, *Algorithms for Clustering Data.* Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.
32. Dhillon, S. Mallela, and D. Modha, “Information-theoretic coclustering,”in *Proc. ACM KDD Conf.*, New York, NY, USA, 2003,pp. 89–98.
33. Banerjee and S. Basu, “Topic models over text streams: A study of batch and online unsupervised learning,” in *Proc. SDM Conf.*, 2007, pp. 437–442.