International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Text Clustering With Using Side Information

Shubhangi V. Airekar[1], **Dhanshree S. Kulkurni**[2]

[1]Department of Computer Engineering, Dr. D. Y. Patil College of Engineerng, Ambi, Pune University, India

[2]Professor, Department of Information Technology, Dr. D. Y. Patil College of Engineerng, Ambi, Pune University, India

Abstract: ***Side information is present along with many text mining application. This side information may be provenance information, any links in the document, web logs which containuser access behavior, the links for any document or any other non textual attributes which are embedded into the text document. All these attributes may contain a large amount of information for clustering purposes. But it is difficult to calculate the concerned importance of this side information especially when some of the data is noisy. In that situation, it is risky to merge side information into the mining process because it can enhance the quality of the representation for the mining process or can add noise in this system. Thus, there should be a proper way to do this mining process so that it will make use of side information to maximize their advantages. Therefore, it is recommended to design an efficient algorithm which makes combination of classical portioning algorithm with probabilistic models in order to create an effective clustering approach.***

Keywords: Data Mining, clustering, text mining, classifier information, text collection.

1.Introduction

The text cluster issue comes in several sort of application domain like the digital information, web, social networks. The space increasing quantity of text information within the encompassing of this massive on-line assortment is the main reason to form economical and climbable mining algorithms. Plenty of work has been done on the problem of cluster in text assortment within the information and data retrival communities. The work is principally designed for the pure text cluster purpose Some example of side-information is given below.

- Web logs contain Meta data which supplies data associated with browsing behavior of varied users. we will track such net logs. Such logs may be accustomed improve the standard of the text mining. Such logs will usually catch sharp interrelatedness in content that cannot be caught by the raw text alone.
- A lot of text documents having connections among them are referred to as attributes. Such links possess plenty of helpful information for mining purpose. Such attributes might typically provide insights concerning the correlation among documents in an exceedingly manner which can be tough to access from raw context.
- Meta information that are unit gift with several net documents might correspond to totally different type of attributes as origin or alternative info concerning the supply of the document. Temporal information, information like possession, location can even be information for mining functions. Documents with user tags additionally return here just in case of network and user sharing application.

Side information can be additional feature for raising the quality of the clustering process but it can be dangerous when the side information is noisy. At that time it can actually degrade the quality of the mining process. Hence an approach is used which gives the combination of the clustering characteristics of the side information and the text content. Thus it is useful in managing the clustering effects in both helpful and noisy data. The main approach of this paper is to determine a clustering where the text attributes and side information give same indications about the character of the original clusters while at the same time ignore aspects in which conflicting indications are provided.

For achieving this goal, portioning approach is merged with probabilistic evaluation method which decides the attachment of the side-information in the clustering process. A probabilisticevaluation process on the side information uses the portioning information for the purpose of evaluation.

2.Related Work

C. C. Aggarwal and P. S. Yu [1] explained the problems of clustering in massive domain data streams where these massive domain data streams are those in which the no. of possible domain values for each attributes are very huge and cannot be easily traced for clustering purposes like IP address streams, credit card transaction stream.

D. Cutting, D. Karger, J. Pedersen, and J. Tukey [2] presented a document viewing method which does document clustering as its primary operation. They also presented fast (linear time) clustering algorithms which helps this interactive browsing paradigm.

M. Steinbach, G. Karypis, and V. Kumar [3] presented experimental study of documents clustering methods by using agglomerative hierarchical clustering and K-means methods .

S. Guha, R. Rastogi, and K. Shim [4] used data with Boolean and categorical attributes for clustering algorithms. A novel concept of links to measure the similarity/proximity between a pair of data points is proposed. This method naturally extend to non-metric similarity measures that are relevant in situation where a domain expert/similarity table is the only source of knowledge.

S. Zhong [5] proposed an efficient online spherical k-means algorithm with an existing scalable clustering strategy to get fast and adaptive clustering of text streams.

Douglass R. Cutting , David R. Karger , Jan O. Pedersen ,John W. Tukey [6] presented document browsing technique that gives document clustering as its main operation is shown here.

A fast clustering algorithms is shown here which support this interactive browsing paradigm.C. C. Aggarwal and C.-X. Zhai [7] discussed the text clustering problems . The key challenges in clustering problems as they are applied to the text are studied .The text clustering methods and their advantages are also shown here. Recent advances in this area of social network and linked data is also discussed here.

The problem of text clustering in context of scalability is studied in [1][9][10].

The problem of text clustering is studied widely by [11][12].

Y. Zhao and G. Karypis [8] presented topic driven clustering where document collection is done according to set of topics. The similarity between documents and topics and relations among documents themselves simultaneously is shown here.

3.Implementation Details

3.1 Description

In text mining application, side data is accessible in the document. Such side data may be contain various types, for example, interfaces in the report, archive provenance data, client access conduct from web logs or other non-literary characteristics. Such properties may contain huge measure of data in the bunching purposes. Nonetheless, the relative data is hard to determine, when some of data is unwanted information. In such cases, it can be unsafe to fuse side-data into the mining procedure, in light of the fact that it can either enhance the nature of the representation for the mining process, or can degrade the performance of the methodology.

In this paper, we outline a calculation which consolidates established apportioning calculations with probabilistic models so as to make a powerful grouping methodology. We then demonstrate to develop the way to the order issue.

3.2. Proposed System

3.2.1 System Description

The proposed system can be summarized in three main steps that are integrated to give accurate results: text document representation, classifier construction and performance evaluation.In the first step, after reading the input text document by the proposed system which divides that text document into features which are also called (tokens, words, terms or attributes), it represents that text document in a vector space as a vector whose components are that features and their weights which are computed by the frequency of each feature in that text document, thereafter it removes the non-informative features (stop words, numbers and special characters). The remaining features are next standardized by reducing them to their root using the stemming process.

3.2.2. System Architecture

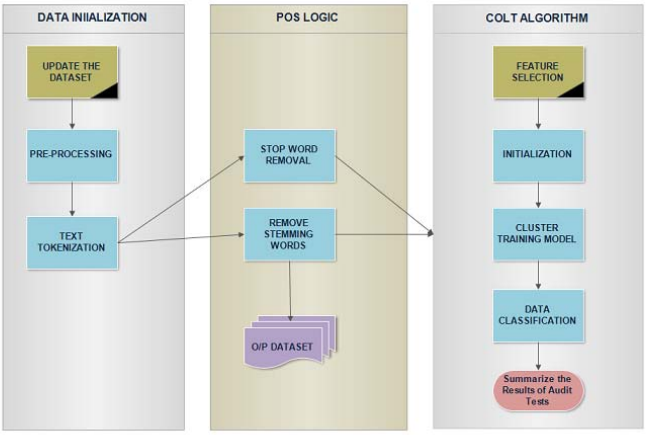


Figure 1: System Architecture

3.2.3 System Modules

- **Text Preprocessing**

Mining from a preprocessed text is easy as compare to natural languages documents. So pre processing of documents that are from different sources is an importantt ask during textmining process before applying any text mining technique. As Text documents can be represented as bag of words on which different textmining methods are based. Let Ω be these to f documents & W={w1,w2,----wm}be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words, which do not provide relevant information; stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed that contain no informatics as such stemming methods: are used to produce the root from the plural or the verbs.

- **Elimination of Stopwords**

In fact, a word which occurs in 80% of the documents in the collection is useless for purposes of retrieval.Such words are frequently referred to as stopwords and are normally filtered out as potential index terms. Articles, prepositions, and conjunctions are natural candidates for a list of stopwords. Elimination of stopwords has an additional important benefit. It reduces the size of the indexing structure considerably. Infact, it is typical to obtain a compression in the size of the indexing structure of 40% or more solely with the elimination of stopwords. A list of 425 stopwords is illustrated. Programs in C for lexical analysis are also provided. Despite these benefit, elimination of stopwords might reduce recall. For instance, consider a user who is looking for documents containing the phrase „to be or not to be.‟ Elimination of stopwords might leave only the term be making it almost impossible to properly recognize the documents which contain the phrase specified.

- **Stemming**

Frequently, the user specifies a word in a query but only a variant of this word is present in a relevant document. This problem can be partially overcome with the substitution of the words by their respective stems. A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes). Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. Many Web search engines do not adopt any stemming algorithm whatsoever. Frakes distinguishes four types of stemming strategies: affix removal, table lookup, successor variety, and n-grams. Table lookup consists simply of looking for the stem of a word in a table. Since such data is not readily available and might require considerable storage space, this type of stemming algorithm might not be practical. Successor variety stemming is based on the determination of morpheme boundaries, uses knowledge from structural linguistics, and more complex than affix removal stemming algorithm.

Colt Algorithm

- *Feature Selection:*
  Feature extraction involves reducing the amount of resources required to describe a large set of data.Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Transforming the input data into the set of features is called feature extraction.
- *Initialization:*
  This step employs a modified k-means approach to initialize clusters, using purely text content, so that each cluster has the records of a particular class only.It involves following steps
- *Cosine Similarity function:*
  Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.

- **Clustering**

Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom - up) and divisive (top - down).An agglomerative clustering starts with one point clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

3.2.3 System Workflow:

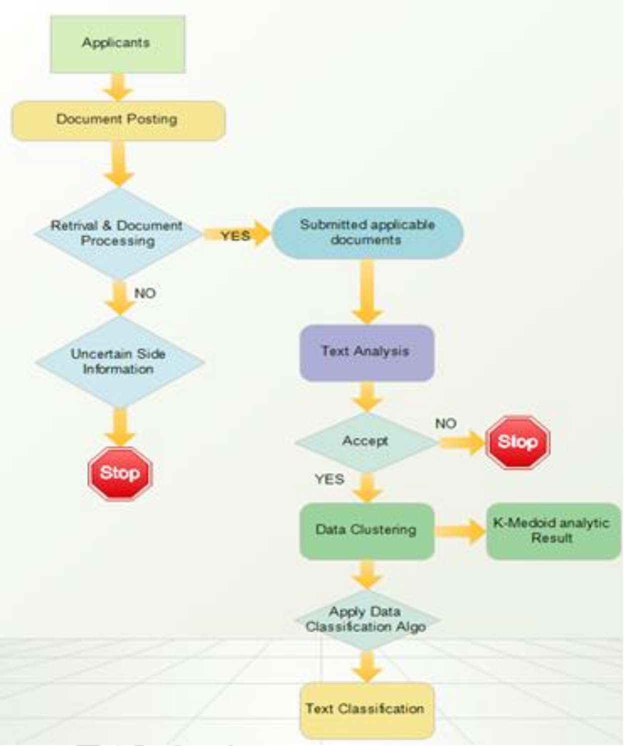
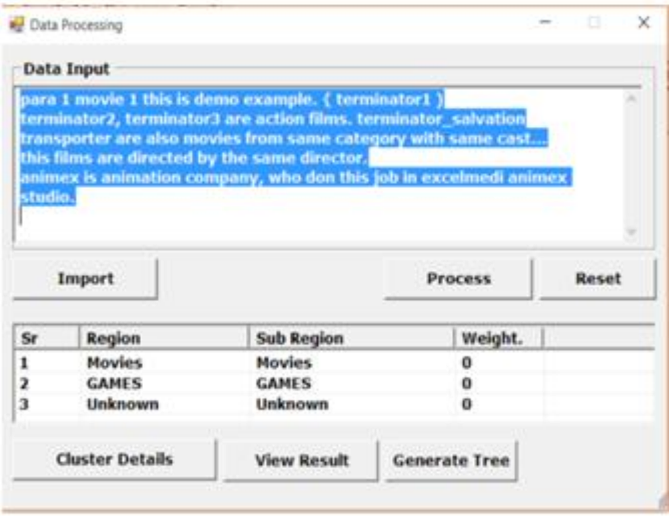


Figure 2: System Workflow

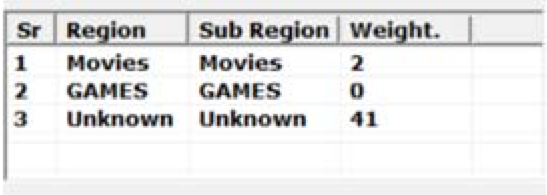
4.Results



para 1 movie 1 this is demo example. (terminator1)
terminator2, terminator3 are action films. terminator_salvation
transporter are also movies from same category with same cast...
this films are directed by the same director.
animex is animation company, who don this job in excelmedi animex
studio.

Sr	Region	Sub Region	Weight.
1	Movies	Movies	0
2	GAMES	GAMES	0
3	Unknown	Unknown	0

Figure 4.1: Data Input



Sr	Region	Sub Region	Weight.
1	Movies	Movies	2
2	GAMES	GAMES	0
3	Unknown	Unknown	41

Figure 4.2: Weight calculation

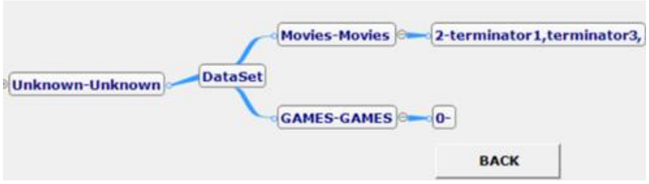


Figure 4.3: Output with side information

Text clustering using side information is shown in above example. When input is given to the system, noisy data is pruned away and remaining side information is compared with the dataset IMDB which is already preprocessed and stored in system. When the side information is found in the dataset, results are created according to it. COATES algorithm gives results which are very close to accuracy. So efficiency and accuracy is more close to result than any other algorithm.

References

[1] C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams," in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[2] 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.

[3] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD,2000, pp. 109–110.

[4] 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.

[5] S. Zhong, "Efficient streaming text clustering," Neural Netw.,vol. 18, no. 5–6, pp. 790–798, 2005.

[6] Douglass R. Cutting , David R. Karger , Jan O. Pedersen , John W. Tukey,"A Cluster-based Approach to Browsing Large Document Collections"

[7] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, NY, USA: Springer, 2012.

[8] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. SIAM Conf. Data Mining, 2005, pp. 358–369.

[9] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in Proc. SDM Conf., 2007, pp. 491–496.999999

[10] S. Zhong, "Efficient streaming text clustering," Neural Netw.,vol. 18, no. 5–6, pp. 790–798, 2005.

[11] 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.

[12] 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.

[13] Gupta, Vishal, and Gurpreet S. Lehal."A survey of text mining techniques and applications." Journal of emerging technologies in web intelligence 1.1 (2009): 60-76.

[14] Megunthadevi, K., et al. "Survey on feature selection in document clustering." International Journal on Computer Science and Engineering 3.3 (2011): 1240-1241

[15] C. C. Aggarwal and P. S. Yu, —On text clustering with side information, in Proc. IEEE ICDE Conf., ashington, DC, USA, 2012.

[16] Preeti Baser, " A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets", 2013

[17] L.V. Bijuraj, "Clustering and its Applications", 2013.