# A Survey – On Side Information for Text Mining

Pranjali Kumbhar*[1], Prof.T.I.Bagban[2]

PG Student, Computer Engineering Department, DKTE, Ichalkarnji, India[1]

Asst. Professor, Information Technology Department, DKTE, Ichalkarnji, India[2]

Pranju.1225@gmail.com[1] , tbagban@yahoo.com[2]

## Abstract-

Any text mining application may contain side information. This side information may be any links in the document, web logs which contain user access behavior, provenance information, the links for any document or any other non-textual attributes which are embedded into the text document. All these attributes may contain a huge amount of information for clustering purposes. But it is difficult to count the concerned importance of this side information especially when some of the data is noisy. In that matter, it is dangerous to merge side-information into the mining process because it can upgrade the quality of the representation for the mining process or can add noise in this system. Thus, there should be a right way to do this mining process so that it will make use of side information to maximize their advantages. Therefore, it is suggested to design an efficient algorithm which makes combination of classical portioning algorithm with probabilistic models in order to create an effective clustering approach. Afterwards, extension to the classification problem is also shown.

Keywords- Text Mining, side Information, Clustering

## 1. Introduction-

Text mining referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. A key element is the linking together of the extracted side information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is dissimilar from what are used to with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is insistent apart all the material that presently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information; rather that no one yet knows and so could not have so far written down.

Text mining usually involves the process of structuring the input text .Usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database, deriving patterns within the structured data, and finally evaluation and interpretation of the output.

Text analysis contains information retrieval, lexical analysis to study word occurrence supplies, pattern recognition, category explanation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The primary aim is, basically, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Text mining is dissimilarity on arena called data mining that tries to discover exciting patterns from huge databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), mentions normally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a new interdisciplinary area which pulls on information retrieval, data mining, machine learning, statistics and computational semantics. As most information is kept as text, text mining is supposed to have a high commercial potential value. Knowledge may be discovered from many foundations of information, however, unstructured texts persist the biggest readily existing source of knowledge. Text mining is analogous to data mining, apart from that data mining tools are intended to switch structured data from databases, but text mining can work with unstructured or semi structured data sets such as emails, full-text documents and HTML files etc. As outcome, text mining is a much improved solution for businesses. To date, still, most research and development efforts have arranged on data mining efforts using structured data. The problem announced by text mining is clear, natural language was developed for humans to link with one another and to record information, and computers are an extended way from understanding natural language.

A lot of work has been done on the issue of clustering in text collection in the database and information retrieval communities. Despite the work is mainly designed for the pure text clustering purpose when other kinds of attributes are absent. Here some example of such side information is given below-

- Web logs contain Meta information which gives information related to browsing behavior of various users. We can track such web logs. Such logs can be used to improve the quality of the text mining. This is because such logs can often catch sharp interrelation in content which cannot be cached by the raw text alone.
- A lot of text documents having connections among them are also called as attributes. Such links possess a lot of useful information for mining purpose. As in the later case, such attributes may often give insights about the correlation among documents in a way which may not be easily accessible from raw context.
- Meta data which are present with many web documents may correspond to different kinds of attributes such as provenance or other information about the source of the document. Temporal information, data such as ownership, location can also be information for mining purposes. Documents with user tags also come here 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 carefully ascertains the coherence of the clustering characteristics of the side information with that of the text content. This helps in managing the clustering effects in both helpful and noisy data. The heart of the approach is to establish a clustering in which the text attributes and side-information provide similar hints about the character of the primary clusters, and at the same time ignore those aspects in which conflicting hints are provided.

To accomplish this target, we will merge a partitioning approach with a probabilistic estimation process, which determines the coherence of the side-attributes in the clustering process. A probabilistic model on the side information uses the partitioning information for the purpose of estimating the coherence of different clusters with side attributes.

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 probabilistic evaluation process on the side information uses the portioning information for the resolution of assessing the attachment of different while our primary goal is to study the clustering problem, we note that such an approach can also be prolonged in principle to additional data mining problems in which auxiliary information is existing with text. This is very common in very wide range of data domains.

## 2. Literature Survey

Paper presented by D. Cutting, D. Karger, J. Pedersen, J. Tukey demonstrates [1] that for information retrieval, document clustering has not been well used. There are two main categories for its objection: first, for large corporation clustering is too slow and second, that retrieval is not improved by clustering. When clustering is used into improve conventional search techniques then only such problems are coming. However, t clustering as an information access tool in its own right obviates these objections, and provides a powerful new access paradigm. Document clustering is presented as primary operation in document browsing technique. Fast clustering algorithms are also presented which support this interactive browsing paradigm.

Paper presented by S. Guha, R. Rastogi, and K. Shim validates [2] that for determining groups and identifying exciting 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 offered which is called CURE that is tougher to outliers, and identifies clusters having non-spherical shapes and extensive variances in size. CURE accomplishes this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them to the center of the cluster by a definite fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. CURE pays a grouping of random sampling and partitioning to handle huge databases. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to gain the desired clusters. In this paper experimental results shows that the quality of clusters produced by CURE is much better than those found by existing algorithms. Further, in this paper it is demonstrated that random sampling and partitioning enable CURE to not only outperforms existing algorithms but also to scale well for large databases without sacrificing clustering quality.

Paper presented by Douglass M. Steinbach, G. Karypis, V. Kumar demonstrates [3] that results of an experimental study of some common document clustering techniques are presented here. In particular, two main approaches to document clustering, agglomerative hierarchical clustering and K-means are compared here. Hierarchical clustering is always the better quality clustering approach, but has limitation due to its quadratic time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters. Sometimes K-means and agglomerative hierarchical approaches are merged so as to ―get the best of both worlds.‖ However, the results shows that the bisecting K-means technique is better than the standard K means approach and as good or better than the hierarchical approaches that we tested for a variety of cluster evaluation metrics. An explanation for these results that is based on an analysis of the specifics of the clustering algorithms and the nature of document data is proposed here.

Paper presented by Charu C. Aggarwal, Stephen C. Gates, and Philip S. Yudem demonstrates [4] that the advantage of using partially supervised clustering is that it is possible to have some control over the range of subjects that one would like the categorization system to address, but with a precise mathematical definition of how each category is defined. An enormously active way then to classify documents is to use this a priori knowledge of the definition of each category. They also discuss a new method to help the classifier distinguish better among closely related clusters.

Paper presented by S. Zhong establishes [5] that clustering data streams has been a novel exploration 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 combines an effective online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means algorithm, using online update based on the well-known. Winner Take all competitive learning. It has been shown to be as effective as SPKM, but much greater in clustering superiority. The scalable clustering strategy was previously developed to deal with very large data bases that cannot fit into a limited memory and that are too expensive to read/scan multiple times. Using this method, one keeps only sufficient statistics for history data to retain the contribution of history data and to accommodate the limited memory. To make the proposed clustering algorithm adaptive to data streams, a forgetting factor is introduced here that applies exponential decay to the importance of history data. The big a set of text documents, the less weight they carry. The experimental results demonstrate the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams—one needs to forget to be adaptive.

Paper presented by C. C. Aggarwal, P. S. Yu demonstrates [6] that real time clustering and segmentation of text data records is required in many applications such as news group filtering, text crawling, and document organization. The categorical data stream clustering problem also has a number of applications to the problems of customer segmentation and real time trend analysis. By making the use of a statistical summarization methodology, an online approach for clustering massive text and categorical data streams is presented here.

Paper presented by C. C. Aggarwal and C.-X. Zhai provides [7] a detailed survey of the problem of text clustering and study the key challenges of the clustering problem, as it applies to the text domain. They discuss the key methods used for text clustering, and their relative advantages also discuss a number of recent advances in the area in the context of social network and linked data.

Paper presented by C. C. Aggarwal and P. S. Yu designs [8] we design an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. They present experimental results on a number of real data sets in order to illustrate the advantages of using such an approach.

## 3. Conclusion

In this paper, approaches are deliberated for mining text data with assembly use of side information. Side information may be obtainable in several procedures of text database which are used to increase the clustering process. Iterative portioning technique is combined with a estimation process to design the clustering method which gives the importance of different kinds of side information. This general method is used to design both clustering and classification algorithms.

## References

[1] 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.

[2] 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.

[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] 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.

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

[6] 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.

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

[8] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.