International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Effective Quality Enhancement of Text Clustering & Classification Using METADATA

Padmaja Shivane[1], Rakesh Rajani[2]

[1]PG Student, Department of Computer Engineering, Alard College of Engineering, University of Pune, Maharashtra, India

[2]Professor, Department of Computer Engineering, Alard College of Engineering, University of Pune, Maharashtra, India

Abstract: ***Text clustering has become more important problem recently because of the large amount of unstructured information which is accessible in many forms in online forums such as the web, online networks, and other information networks. In a lot of cases, the information is not purely available in text form. A lot of side-information is available along with the text documents. Such side-information may be of altered kinds, such as the links in the document, user-access behaviour from web logs, or added non-textual attributes which are embedded into the text document. Such attributes may contain a large amount of data for clustering purposes. However, the data relativity of this side-information may be difficult to estimate, abnormally if some of the information is noisy. In such cases, it can be chancy to absorb side information into the clustering technique, because it can either improve the superior of the representation for clustering, or can add noise to the process. Therefore, we charge a conscionable way to perform the clustering technique, so as to aerate the advantages from application this side information. In this paper, we survey on side information for improving the text mining technique.***

Keywords: Text clustering, side-information, text mining, clustering technique.

1. Introduction

The issue of text clustering arises in the ambience of many application domains such as the web, networks, and other digital collections. The rapidly accretion amounts of text data in the ambience of these huge online collections has led to an absorption in creating scalable and effective mining algorithms. A huge amount of plan has been done in recent years on the botheration of clustering in data collections [1] [2] [3] [4] [5] in the database and retrieval communities. However, this plan is primarily advised for the botheration of authentic text clustering, in the absence of other kinds of attributes. In abounding appliance domains, a tremendous amount of side-information is as well associated forth with the documents. This is because argument abstracts about action in the ambience of an array of applications in which there may be a ample bulk of added kinds of database attributes or Meta information which may be advantageous to the absorption process. Some examples of such side-information are as follows:

1. In an application within which we tend to track user access behavior of web documents, the user-access behavior is also stored within the web logs. For every document, the meta-information could relate to the searching behavior of the various users. Logs like these may be utilized for improving the standard of the mining method in a very method that is a lot of meaningful to the user, and conjointly application-sensitive. Typically this will be as a result of the logs can frequently select refined correlations in content that cannot be selected by the raw text alone.
2. Several text documents have links among them, which might even be treated as attributes. Such links enclose lots of helpful data for mining functions. As within the previous case, such attributes could usually give insights concerning the correlations among documents in a very method which cannot be simply accessible from raw content.
3. Several web documents have meta-data related to them that correspond to totally various sorts of attributes like the place of origin or different data concerning the origin of the document. In different cases, knowledge like ownership, location, or perhaps temporal data is also informative for mining functions. In a very variety of network and user-sharing applications, documents are also related to user-tags, which can even be quite informative.

While such side-information will typically be helpful in enhancing the standard of the clustering method, it will be a dangerous method once the side-information is noisy. In such situations, it will truly worsen the standard of the mining method. Thus, associate approach is utilized that precisely determines the coherence of the cluster characteristics of the side information there upon of the text content.

This is useful in amplifying the clustering effects of each variety of data. The fundamentals of the approach is to work out a clustering which the text attributes and side-information give similar hints regarding the nature of the underlying clusters, and at identical time ignore those aspects during which conflicting hints are given. In order to attain this goal, a partitioning approach can be combined with a probabilistic estimation method that determines the coherence of the side-attributes within the cluster method. A probabilistic model on the side information utilizes the partitioning information from text attributes with the aim of estimating the coherence of various clusters with aspect attributes. This is useful in extracting out the noise within the membership behavior of various attributes. The partitioning method is specifically designed to be highly economical for substantial datasets. This could be vital in situations during which the datasets are substantial.

2. Literature Review

C. C. Aggarwal and H. Wang [6] conferred a survey of graph mining and management applications. They additionally gave a survey of the common applications that arise within the context of graph mining applications. Most of the research in recent years has targeted on little and memory-resident graphs. Numbers of the long run challenges arise within the context of substantial disk-resident graphs. Alternative vital applications area unit designed within the context of large graphs streams. Graph streams arise within the setting of variety of applications like social networking, within which the communications between huge groups of users area unit captured within a graph. Such applications are difficult, since the whole data cannot be localized on disk for the aim of structural analysis. Therefore, new techniques are needed to summarize the structural behavior of graph streams, and use them for a range of analytical situations.

The classification issue is one amongst the foremost elementary issues within the machine learning and data processing literature [7]. Within the context of text data, the issue may also be thought of kind of like that of classification of distinct set-valued attributes, once the frequencies of the words square measure overlooked. The domains of those sets rather huge as it contain the complete lexicon. Therefore, text mining methods have to be compelled to be designed to effectively manage huge numbers of parts with varied frequencies. Most of the popular methods for classification like decision trees, rules, Bayes technique, nearest neighbor classifiers, SVM classifiers, and neural networks are extended to the case of text data. Recently, a substantial quantity of stress has been placed on linear classifiers like neural networks and SVM classifiers, with the latter being notably suited to the characteristics of text data. In last few years, the advancement of internet and social network technologies have result in an amazing interest within the classification of text documents consisting links or meta-information. Recent analysis has shown that the incorporation of linkage info into the classification method will considerably improve the standard of the underlying results.

C. C. Aggarwal and P. S. Yu [8] proposed a technique for text clustering with the utilization of side-information. Several types of text databases contain a substantial quantity of side-information or meta-information, which can be employed in order to enhance the clustering method. So as to design the clump methodology, they combined a reiterative partitioning method with a probability estimation method that computes the significance of various forms of side-information. They proposed results on real datasets illustrating the effectiveness of their methodology. The results demonstrate that the utilization of side-information will greatly improve the standard of text clustering, whereas maintaining a high level of effectiveness.

J. Chang and D. Blei [9] developed the Relational Topic Model (RTM), a model of documents and therefore the links between them. For every pair of documents, the RTM models their link as a binary variable which is conditioned on their contents. The model is utilized for summarizing a network of documents, predict links among them, and predict words inside them. They had a tendency to derive economical illation and learning algorithms supported variational strategies and measure the prognostic performance of the RTM for substantial networks of scientific abstracts and web documents. The RTM is a novel probabilistic generative model of documents and links among them. The RTM is employed to research coupled corpora like citation networks, coupled web content, and social networks. They showed qualitatively and quantitatively that the RTM gives an efficient and helpful technique for analyzing and utilizing this data. It considerably enhances on existing models, combining each node-specific information and link structure to provide higher predictions.

In the study by Y. Sun et. al [10], a new technique of generative topic model known as iTopicModel is presented that integrates each network structure and text data for the most commonly found document networks. This model features two-layered graphical model architecture. First, a practical variable Markov Random Field is characterized to model the dependency relations between documents. Second, a conventional document generative model is employed, that is not absolutely autonomous with one another given this topic distribution configurations. A joint probability method is then characterized depended on the graphical model. They presented associate EM-based iterative resolution to calculate the set of best parameters that empowers the log-likelihood of the joint distribution. Their experiments demonstrate that this model is more successful than the progressive topic modeling ways, in each aspect: following human perception and being reliable with the network configuration. Also, they demonstrate that presented model, with the assistance of Q-function, will facilitate us mechanically construct concept hierarchy in on-line databases.

Y. Zhou et. al [11] had solved the issue of graph clustering depended structural and attribute similarities. A unified neighborhood stochastic process distance measure is built to calculate vertex closeness on associate degree attribute increased graph. Depended on this distance measure, they have a tendency to take a *K*-Medoids clump approach to separate the graph into *k* clusters that have each cohesive intra-cluster structures and homogenized attribute values. They offer hypothetical analysis to quantitatively calculate the contributions of attribute similarity within the stochastic process distance measure. They have built a learning algorithmic program to regulate the degree of contributions of various attributes within the stochastic process model as they have a tendency to iteratively refine the clusters, and prove that the weights are accustomed towards the direction of clustering convergence.

Q. Mei [12] had met a knowledge assortment with substantial textual information and a network structure in several data discovery tasks. Statistical topic models extricate coherent topics from the text, whereas typically overlooking the network structure. Social network analysis on the opposite hand has a tendency to target the topological network structure, whereas keeping aside the textual data. They have a tendency to formally characterize the most

important tasks of topic modeling with network structure. They presented a general answer of text mining with network structure that optimizes the probability of topic generation and therefore the topic smoothness on the graph in a much unified means.

Specifically, they present a regularization technique for statistical topic models, with a harmonic regularizer depended on the network structure. The final technique permits impulsive selections of the topic model and therefore the graph based regularizer. They demonstrated that with concrete selections, the model is applied to handle real world text mining issues like author-topic analysis, topical community discovery, and spatial topic analysis. Practical experiments on two entirely different genres of knowledge demonstrate that their projected methodology is efficient to extract topics, find topical communities, construct topic maps, and model geographic topic distributions. It enhances each pure topic modeling, and pure graph-based methodology.

3. Proposed System

The primary objective of this paper is to check the clustering issue. This type of approaches may be extended in essence to alternative data mining issues within which auxiliary information is accessible with text. Such situations are quite common in large sorts of knowledge domains. Thus, a method is proposed to increase the approach to the problem classification. The expansion of the approach to the classification problem gives superior results as a result of the incorporation of side-information. The goal is to indicate that the benefits of utilizing side-information extend on the far side of a pure clustering task, and might give competitive benefits for a wider form of problem situations.

4. Conclusion

Techniques for mining text information with the employment of side-information are presented. Several types of text-databases consist of substantial quantity of side-information or meta-information, which can be utilized in order to enhance the clustering method. To construct the clustering methodology, reiterative partitioning technique is combined with a probability estimation method that computes the significance of various types of side-information. This general method is utilized to build each clustering and classification algorithms.

Reference

[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.
[2] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections,” in ACM SIGIR Conf.
[3] H. Schutze and C. Silverstein, “Projections for Efficient Document Clustering,” in ACM SIGIR Conf., 1997.
[4] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in Text Mining Workshop, KDD, 2000.
[5] S. Zhong, “Efficient Streaming Text Clustering,” Neural Networks, vol 18, 2005.
[6] C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*. New York, NY, USA: Springer, 2010.
[7] C. C. Aggarwal and C.-X. Zhai, “A survey of text classification algorithms,” in 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.
[9] J. Chang and D. Blei, “Relational topic models for document networks,” in Proc. AISTASIS, Clearwater, FL, USA, 2009.
[10] Y. Sun, J. Han, J. Gao, and Y. Yu, “iTopicModel: Information network integrated topic modeling,” in Proc. ICDM Conf., Miami, FL, USA, 2009.
[11] Y. Zhou, H. Cheng, and J. X. Yu, “Graph clustering based on structural/attribute similarities,” PVLDB, vol. 2, no. 1,2009.
[12] Q. Mei, D. Cai, D. Zhang, and C.-X. Zhai, “Topic modeling with network regularization,” in Proc. WWW Conf., New York, NY, USA, 2008, pp. 101–110.

Author Profile

Padmaja Shivane is PG Student of Department of Computer Engineering, at Alard College of engineering and management, University of pune.

Professor Rakesh Rajani is in Department of Computer Engineering at Alard College of Engineering and Management, University of Pune.