Survey Paper on Descriptive Clustering of Documents on the Basis of Predictive Network

Viral Vala[1], Prof. Abhijit Patankar[2]
[1]*PG Student*
[2]*Faculty*
[3]*School of Computer Engineering Technology,*
Alard College of Engineering.

Abstract– every day the mass of information available, merely finding the relevant information is not the only task of automatic text classification systems. Instead the automatic text classification systems are assumed to retrieve the relevant information and are organized according to their degree of relevance to the query. The main problem in organizing is to classify which documents are relevant and which are irrelevant. The Automated text classification consists of automatically organizing clustered data. In this survey we learned method of text classification using machine learning based on the bag of words. The closest ancestors of the senses of all the undamaged words in a given document are selected as classes for the specified document.

Keywords- Descriptive clustering, feature selection, K-means clustering, model selection, machine learning.

I. Introduction

Every day the mass of information available to us increases. This information would be irrelevant if our ability to productively get to did not increment too. For most extreme advantage, there is need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising region is the automatic text categorization. Envision ourselves within the sight of impressive number of texts, which are all the more effectively available on the off chance that they are composed into classes as per their topic. Obviously one could request that human read the text and arrange them physically. This assignments is hard if done on hundreds, even a huge number of texts. Thus, it appears to be important to have a computerized application, so here automatic text categorization is presented. An expanding number of information mining applications includes the investigation of unpredictable and organised information types and requires the utilisation of expressive example dialects. Many of these applications cannot be solved using traditional data mining algorithms. This observation is the main motivation of Clustering.

Unfortunately, existing “upgrading" approaches, especially those that use logical programming techniques, often suffer not only from poor scalability when it comes to complex database schemas, but also from unsatisfactory predictive performance when managing numeric or noisy values. In real-world applications. However, "flattening" strategies tend to take a lot of time and effort to transform data, result in the loss of compact representations of standardized databases and produce an extremely large table with a large number of additional attributes and numerous NULL values (lost values). As a result, these difficulties have prevented wider application of multi-relational mining and represent an urgent challenge for the mining community. To address the above mentioned problems, this paper introduces a Descriptive clustering approach where neither “upgrading” nor “flattening” is required to bridge the gap between propositional learning algorithms and relational.

In Proposed approach, Data analysis techniques, such as clustering it can be used to identify subsets of data instances with common characteristics. Users can explore the data by examining some instances in each group instead of rather than examining the instances of the complete data set. This allows users to focus efficiently on large relevant subsets Data sets, in particular for document collections. In particular, the descriptive grouping consists of automatic grouping sets of similar instances in clusters and automatically generate a description or a synthesis that can be interpreted by man for each group. The description of each cluster allows a user determine the relevance of the group without having to examine its content For text documents, a description suitable for each group can be a multi-word tag, an extracted title or a list of characteristic words . The quality of the grouping it is important, so that it is aligned with the idea of likeness of the user, but it is equally important to provide a user with a brief and informative summary that accurately reflects the contents of the cluster

A. *Motivation*

In introductory part for the study of Text Classification, their application, which algorithm used for that and the different types of model, I decided to work on the Text Classification which is used for data analysis lot of work done on that application and that the technique used for that application is Text Classification using traditional data mining algorithms and not worked on word sense technique.

Approaches to the state of the art to classify data it can be used to identify subset of data instances. However, they suffer from low accuracy.

II. Related Work

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

In this section, we briefly review the related work on Text classification and their different techniques.

J.-T. Chien, describe the “Hierarchical theme and topic modeling,” in that Taking into account hierarchical data sets in the body of text, such as words, phrases and documents, they perform structural learning and deduce latent themes and themes for sentences and words from a collection of documents, respectively. The relationship between arguments and arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. A tree branching process is presented to draw the proportions of the topic for different phrases. They build a hierarchical theme and a thematic model, which flexibly represents heterogeneous documents using non-parametric Bayesian parameters. The thematic phrases and the thematic words are extracted [1].

Bernardini, C. Carpineto, and M. D'Amico, describe the "Full-subtopic retrieval with keyphrase-based search results clustering," in that Consider the problem of restoring multiple documents that are relevant to the individual sub-topics of a given Web query, called "full child retrieval". To solve this problem, they present a new algorithm for grouping search results that generates clusters labeled with key phrases. The key phrases are extracted generalized suffix tree created by the search results and merge through a hierarchical agglomeration procedure improved grouping [2].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, describe the "Self-organization of a massive document collection," this paper describes the implementation of a system that can organize large collections of documents based on textual similarities. It is based upon the self-organized map (SOM) algorithm. Like the feature vectors for documents, the factual portrayals of their vocabularies are utilized. The main objective of our work was to resize the SOM algorithm in order to handle large amounts of high-dimensional data [3].

K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, describe the "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," in that Organizing Web search results in a hierarchy of topics and secondary topics makes it easy to explore the collection and position the results of interest. In this paper, they propose a new hierarchical monarchic grouping algorithm to construct a hierarchy of topics for a collection of search results retrieved in response to a query. At all levels of the hierarchy, the new algorithm progressively identifies problems in order to maximize coverage and maintain the distinctiveness of the topics. They refer to the algorithm proposed as DisCover. The evaluation of the quality of a hierarchy of subjects is not a trivial task, the last test is the user's judgment. They have used various objective measures, such as coverage and application time for an empirical comparison of the proposed algorithm with two other monotetic grouping algorithms to demonstrate its superiority [4].

R. Xu and D. Wunsch, describe the "Survey of clustering algorithms," in that Information investigation assumes an irreplaceable job in understanding the different marvels. Combination investigation, crude investigation with almost no past information, comprises of research created in a wide assortment of networks. Assorted variety, from one perspective, furnishes us with numerous devices. Then again, the abundance of choices causes perplexity. They have analyzed the gathering calculations for the informational indexes that show up in measurements, software engineering and AI and they represent their applications in some reference datasets, the issue of road merchants and bioinformatics, and another field that draws in exceptional endeavors [5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami, describe the "Inductive learning algorithms and representations for text categorization," in that Information investigation assumes an irreplaceable job in understanding the different marvels. Combination investigation, crude investigation with almost no past information, comprises of research created in a wide assortment of networks. Assorted variety, from one perspective, furnishes us with numerous devices. Then again, the abundance of choices causes perplexity. They have analyzed the gathering calculations for the informational indexes that show up in measurements, software engineering and AI and they represent their applications in some reference datasets, the issue of road merchants and bioinformatics, and another field that draws in exceptional endeavors [6].

R. Kohavi and G. H. John, describe the "Wrappers for feature subset selection, "In that the highlight subset choice issue, a learning calculation is looked with the issue of choosing a pertinent subset of highlights whereupon to concentrate, while overlooking the rest. To accomplish the most ideal execution with a specific learning calculation on a specific preparing set, a component subset determination strategy ought to think about how the calculation and the preparation set collaborate. They investigate the connection between ideal component subset choice and significance. Our wrapper technique looks for an ideal element subset custom-made to a specific calculation and a space. They contemplate the qualities and shortcomings of the wrapper approach and demonstrate a progression of enhanced plans. They contrast the wrapper approach with enlistment without highlight subset choice and to Relief, a channel way to deal with highlight subset choice. Critical enhancement in precision is accomplished for some datasets for the two groups of enlistment calculations utilized: choice trees and Naive-Bayes [7].

T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, describe the "Self-organization of a massive document collection," This paper describes the implementation of a system that is able to organize vast document collections according to textual similarities. It is based on the self-organizing map (SOM) algorithm. As the feature vectors for the documents statistical representations of their vocabularies are used. The main goal in our work has been to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data [8].

Q. Mei, X. Shen, and C. Zhai, describe the "Automatic labeling of multinomial topic models," In this paper, they propose probabilistic ways to deal with consequently marking multinomial point models in a goal way. They give this marking issue a role as an improvement issue including limiting Kullback-Leibler difference between word circulations and expanding common data between a name and a theme display. Analyses with client examine have been done on two content informational collections with various classes. The outcomes demonstrate that the proposed marking techniques are very compelling to create names that are significant and valuable for deciphering the found point models [9].

K. Lagus and S. Kaski, describe the "Keyword selection method for characterizing text document maps," in that Characterization of subsets of data is a recurring problem in data mining. They propose a keyword selection method that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated, with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to characterizing graphical displays of collections of data ordered e.g. with the SOM algorithm [10].

III. Open Issue

Lot of work has been done in this field because of its extensive usage and applications. In this section, some of the approaches which have been implemented to achieve the same purpose are mentioned. These works are majorly differentiated by the algorithm for Text Classification.

In another research, to access the relevant information from mass of data is very difficult and time consuming task as every day mass of information increases because of digital world. Every day, the mass of information available to us increases. This information would be irrelevant if our ability to efficiently access did not increase as well. Automated text classification provide us with maximum benefit that allow us to search, sort, index, store, and analyze the available data. It also allows us to find in desired information in a reasonable time.

As my point of view when I studied the papers the issues are related to Text Classification. The challenge is to addressing automatic text classification problem using machine learning.

Conclusion

Proposed descriptive Clustering as two coupled predictions activity choose a grouping that is predictive of features and prediction of the cluster assignment of a subset of features. Use predictive performance as a goal criterion, descriptive clustering parameters the number of clusters and the number of functions per Clusters: they are chosen from the model selection. With the result solution, each group is described by a minimum subset of features necessary to predict if an instance belongs to the cluster our hypothesis is that even a user will be able to predict membership in the group of documents using the descriptive features selected by the algorithm. Given Some relevant requirements, a user can quickly identify clusters that probably contain relevant documents

References

[1] J.-T. Chien, “Hierarchical theme and topic modeling,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565–578, 2016.

[2] Bernardini, C. Carpineto, and M. D’Amico, “Full-subtopic retrieval with keyphrase-based search results clustering,” in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206–213.

[3] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, “Self-organization of a massive document collection,”IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[4] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram, “A hierarchical monothetic document clustering algorithm for summarization and browsing search results,” in Proc. Int. Conf. World Wide Web, 2004, pp. 658–665.

[5] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.

[6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” in Proc. Int. Conf. Inform. Knowl. Manag., 1998, pp. 148–155.

[7] R. Kohavi and G. H. John, “Wrappers for feature subset selection,”Artif. Intell., vol. 97, no. 1, pp. 273–324, 1997.

[8] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero,and A. Saarela, “Self-organization of a massive document collection,”IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 574–585, 2000.

[9] Q. Mei, X. Shen, and C. Zhai, “Automatic labeling of multinomial topic models,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2007, pp. 490–499.

[10] K. Lagus and S. Kaski, “Keyword selection method for characterizing text document maps,” in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp. 371–376.

[11] Patankar, A.J., Sirbi, K, Kulhalli and K.V, “Preservation of privacy using multidimensional k-anonymity method for non-relational data Open Access” in Int. Journal of Recent Technology and Engineering (IJRTE), 2019, pp. 371–376.

[12] Patankar, A.J., Sirbi, K, Kulhalli and K.V, “Emotweet: Sentiment Analysis tool for twitter” in IEEE Int. Conference on Advances in Electrical Communication and Computer Technology (ICAECCT), 2019, pp. 371–376.