Survey Paper on Self-Tuned Descriptive Document Clustering using a Predictive Network

Prof.S.N.Shelke, Assistant Professor,Department of Computer Engg,Sinhgad Academy of Engineering,
Students : Balaji C. Narwade , Akash D.Surwase , Sahil S.Zele , Bhalchandra A. Patil
Sinhgad Academy of Engineering, Department of Computer Engineering , Savitribai Phule Pune University

Abstract—the descriptive grouping consists of automatically organizing data instances into groups and generating a description summary for each group. The description should inform a user about the content of each group without further examination of the specific instances, allowing a user to quickly scan relevant groups. Selection of descriptions is often based on heuristic criteria. We modeling the descriptive cluster as an auto-coder network that predicts the characteristics of cluster assignments and predicts the cluster assignments of a subset of characteristics. The subset of functionality used to predict a cluster serves as a description. For the text documents, appearance or counting of words, phrases or other attributes provides a representation of the dispersed features with interpretable feature labels In the proposed network, cluster predictions are performed using logistic regression models and feature forecasts are based on logistic regression models. The optimization of these models leads to a completely self-regulating descriptive grouping approach that automatically selects the number of clusters and the number of functions for each cluster. We apply the methodology to a variety of short text documents and has shown that the selected grouping, as demonstrated by the subsets of the selected features, is associated with a significant topical organization.

Keywords- Descriptive clustering, feature selection, logistic regression, model selection.

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 a need of devices that permit look, sort, list, store and investigate the accessible information. One of the promising regions is the automatic text categorization. Envision ourselves within the sight of an 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 assignment 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 increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the Linear Regression.

Unfortunately, existing "upgrading" approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, "flattening" strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with a huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi-relational mining, and post an urgent challenge to the data mining community. To address the above-mentioned problems, this article 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 Datasets, 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 a 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

II. RELATED WORK

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. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely "look for secondary arguments length under the sufficiency of k documents". they have used a test collection specifically designed to evaluate the recovery of the sub-themes, they have found that our algorithm has passed both other clustering algorithms of existing research results as a method of redirecting search results underline the diversity of results (at least for k> 1, that is when they are interested in recovering more than one relevant document by sub-theme) [1].

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 monothetic grouping algorithms to demonstrate its superiority. Although our algorithm is a bit more computationally than one of the algorithms, it generates better hierarchies. Our user studies also show that the proposed algorithm is superior to other algorithms as a tool for summary and navigation [2].

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, we perform structural learning and we 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. In the experiments, the proposed method is evaluated as effective for the construction of a semantic tree structure for the corresponding sentences and words. The superiority of the use of the tree model for the selection of expressive phrases for the summary of documents is illustrated [3].

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 on the self-organized map (SOM) algorithm. Like the feature vectors for documents, the statistical representations of their vocabularies are used. The main objective of our work was to resize the SOM algorithm in order to handle large amounts of high-dimensional data. In a practical experiment, they mapped 6 840 568 patent abstracts in a SOM of 1.002.240 nodes. As characteristic vectors, we use vectors of 500 stochastic figures obtained as random projections of histograms of weighted words [5].

Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu, describe the "Using WordNet to Disambiguate Word Senses for Text Classification," in that they propose an automatic method of text classification. Based on the disambiguation of the meaning of words. We use the "bell" algorithm to eliminate the word ambiguity so that every word is replaced by its meaning in context. The closest ancestors of the senses of all words without stopping in a given document Selected as classes for the specified document [5].

S. Dumais, J. Platt, D. Heckerman, and M. Sahami, describe the "Inductive learning algorithms and representations for text categorization," in that Text categorization the assignment of natural language texts to one or more predefined categories based on their content is an important component in many information organization and management tasks. They compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed and classification accuracy. They also examine training set size, and alternative document representations. Very accurate text classifiers can be learned automatically from training examples. Linear Support Vector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate [6].

III. EXISTING SYSTEM

In another research, to access the relevant information from a mass of data is a very difficult and time-consuming task as an everyday mass of information increases because of a 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 provides us with maximum benefit that allows 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 regression.

IV. PROPOSED APPROACHES

In Proposed System training is a creation of train dataset using which classification of unknown data in predefined categories is done. Here a learning system is created using regression. It is a supervised learning where unlabeled data is classified using labeled data. Training data is always a labeled dataset based on its features.

The project had considered no scientific papers form different publication of different domains for creating training dataset. These papers are input for creating training dataset. This input is first preprocessed and most informative features are extracted using TF/IDF algorithm. Ten different domains from the market are identified and then extracted feature and have to put to a corresponding domain where each domain is considered as one class that which is used for labeling test dataset in testing part and features are considered as nodes. Once the training part is completed, all features of respective domains are get updated in corresponding tables in the database.

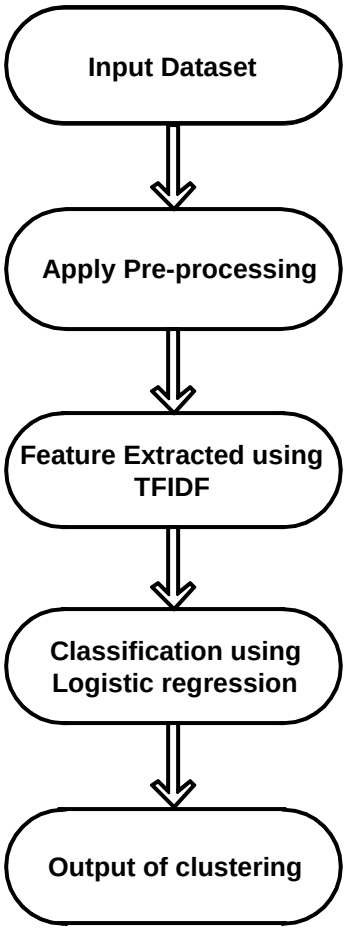

Fig. Flow Diagram

V. CONCLUSION

Proposed Text classification as two coupled predictions activity chooses a grouping that is predictive of features. Use predictive performance as a goal criterion, classification parameters the number of function: 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 data our hypothesis is that even a user will be able to predict membership in the group of documents using the features selected by TFIDF and the classification using logistic regression. Given Some relevant requirements, a user can quickly identify that probably contain relevant documents

VI. 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] Ying Liu1, Peter Scheuermann2, Xingsen Li1, and Xingquan Zhu1 Using WordNet to Disambiguate Word Senses for TextClassification.

[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.