ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

A Review on Document Clustering Using a Machine Learning Framework

Sidhesh Sawalkar, Sahil Swapnil, Rajat Dave, Rudransh Sinha, Vaishali Wangikar

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

School of Computer Engineering, MIT Academy of Engineering, Pune, Alandi, Maharashtra, India

ABSTRACT: The descriptive grouping consists of automatically organizing data instances into groups and generating a descrip-tion 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 modelling 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 and feature forecasts are based on machine learning framework. 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 text documents and has shown that the selected grouping, as demonstrated by the subsets of the selected features, is associated with significant topical organization.

KEYWORDS: Self-tuned, Descriptive clustering, feature selec-tion, machine learning.

I. INTRODUCTION

Every day the mass of information available to us increases. This information would be irrelevant if the ability of produc-tivity does not increase. 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. One could request a human to read the text and arrange them physically. These assignments are hard if done on huge number of texts. Thus, it appears to be important to have a computerized application, so automatic text categorization is introduced. An increasing amount of data mining applications involve the analysis of complex and structured types of data and requires 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 using the machine learning approach.

Existing upgrading approaches, especially those using Logic Programming techniques, often suffer poor scalability while dealing with complex database schemas but also an unsatis-factory predictive performance in handling noisy or numeric values in the real world applications. However, flattening strategies tend to require considerable time and effort for the data transformation, result in losing the compact representa-tions of the normalized databases, and produce an extremely large table with huge number of additional attributes and missing values. As a result, these

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

difficulties have prevented a wide application of multi-relational mining, and post a number of challenges 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 and relational learning algorithms.

In Proposed approach such as clustering can be used to identify subsets of data instances with common characteristics. Users can explore the data by examining some instances in each group rather than considering the instances of the complete data set. This allows users to focus efficiently on a large relevant subset of datasets for document collections. Particularly, 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 to 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 is important so that it aligns with the interest of the user, but it is equally important to provide a user with a brief and informative summary which accurately reflects the contents of the cluster.

Some of the applications of document clustering are

1. Automatic document organization

2. Topic extraction

3. Fast information retrieval

4. Analysis of customer feedback

5. Discovering meaningful implicit subjects across documents 6.Optimization of web search engine result

II. RELATED WORK

In paper [1],Kohonen et al. 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. As the feature vectors for documents, the statistical representations of their vocabularies are used. The main objective of their work was to resize the SOM algo-rithm in order to handle large amounts of high-dimensional data. In this paper[2] taking into account hierarchical data sets in the body of text, such as words, phrases and docu-ments, Chien et al. perform structural learning and deduce latent themes for sentences and words from a collection of documents, respectively. The relationship between arguments in different data groupings is explored through an unsupervised procedure without limiting the number of clusters. In paper [3] Bernardini et al. 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 from generalized suffix tree created by the search results and merge through a hierarchical ag-glomeration procedure giving improved grouping. They also introduce a new measure to evaluate the performance of full recovery sub-themes, namely looking for secondary arguments length under the sufficiency of k documents. In paper [4] Singal et al. organize web search results in a hierarchy of topics and secondary topics making it easy to explore the collection and position 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 monotonic grouping algorithms to demonstrate its superiority. Although their algorithm is a bit more heavy

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

computationally than one of the other algorithms, it generates better hierarchies. In paper [5] of Xu et al. data analysis plays an indispensable role in understanding the various phenomena. Conglomerate analysis, primitive exploration with little or no previous knowledge, consists of research developed in a wide variety of communities. They have examined the grouping algorithms for the data sets that appear in statistics, computer science and machine learning and they illustrate their applications in some reference datasets.Dumais et al. consider text categorization as the assignment of natural lan-guage texts to one or more predefined categories based on their important content in many information organization and management tasks[6]. 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 exam-ples. Linear Support Vector Machines (SVM) are particularly promising because they are very accurate, quick to train and quick to evaluate. In paper[7],Kohavi et al. consider the feature subset selection problem.Learning algorithm is faced with the problem of selecting a relevant subset of features upon which to focus its attention, while ignoring the rest. To achieve the best possible performance with a particular learning algorithm on a particular training set, feature subset selection method should consider how the algorithm and the training set interact. They explore the relation between optimal feature subset selection and relevance.They use a wrapper method which searches for an optimal feature subset tailored to a particular algorithm and a domain. They study the strengths and weaknesses of the Wrapper approach and show a series of improved designs. This paper [8] 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. The main goal of their work was to scale up the SOM algorithm to be able to deal with large amounts of high-dimensional data. In a practical experiment, they mapped 6,840,568 patent abstracts onto a 1,002,240-node SOM. As feature Vectors, they used 500-dimensional vectors of stochastic figures obtained as random projections of weighted word histograms.Mei et al. propose probabilistic approaches to automatically label multinomial topic models in an objective way in paper[9]. They cast the labelling problem as an optimization problem involving minimizing Kullback-Leibler divergence between word distributions and maximizing mutual information between a label and a topic model.They did experiments with user study on two text data sets with different genres. The results showed that the proposed la-belling methods were quite effective to generate labels that are meaningful and useful for interpreting the discovered topic models. Their methods are general and can be applied to labelling topics learned through all kinds of topic models such as PLSA, LDA, and their variations. In paper [10],Lagus et al.consider the recurring problem of Characterization of subsets of data.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. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map.

III. EXISTING SYSTEM

There has been extensive research on clustering and other unsupervised methods, namely topic modeling but also low-dimensional embedding for exploring text datasets. Some particularly relevant work has focused on website search results. We highlight approaches that have considered user interpretation of the results in the form of descriptive keywords, phrases, or titles that summarize the semantic content of clusters and topics for a user. For datasets with known ground-truth topic categories, document clustering can be evaluated using correspondence measures, but comparisons of descriptive labels have often relied on human evaluation. Comparisons among description mechanisms is especially challenging, since the datasets, clustering or modeling paradigms, and form of the description varies widely. Although some user evaluations have concentrated on which labels users prefer, evaluation should concentrate on whether the descriptions aid of a user in predicting the most relevant cluster or topic. To our knowledge, no previous approach has posed descriptive clustering as a prediction problem with objective quantification in terms of classification performance. The other unique contributions of our approach include a principled approach to select the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

number of features in the description, and an automatic approach for selecting the number of clusters that is independent of the clustering algorithm but dependent on the cluster assignments and the binary feature representation.

After literature review following issues are not addressed by the existing system

1 Not able to worked on automated text classification.

2 Domain wise analysis is time consuming

3. Feature subsets were not selected automatically

4. There was no principled approach to feature selection

IV. PROBLEM STATEMENT

To access the relevant information from huge amount of data is very difficult and time consuming task as every day mass of information increases due to digital world. It is proposed automated text categorization application which helps grouping of logical distributions similar to pdf with increased accuracy and reduced time complexity.

A. PROPOSED APPROACH

In Proposed System, training data set is used which clas-sifies the unknown data in predefined categories.A supervised learning system uses machine learning algorithms where un-labelled data is classified into labelled data. Training data is always a labelled dataset based on its features.

Project had considered number of scientific papers from 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 market are identified and then extracts its features to allocate its corresponding domain where each domain is considered as one class that is used for labelling the test dataset in testing part and its features are considered as nodes. Once training part is completed, all features of respective domains get updated in corresponding tables in database.

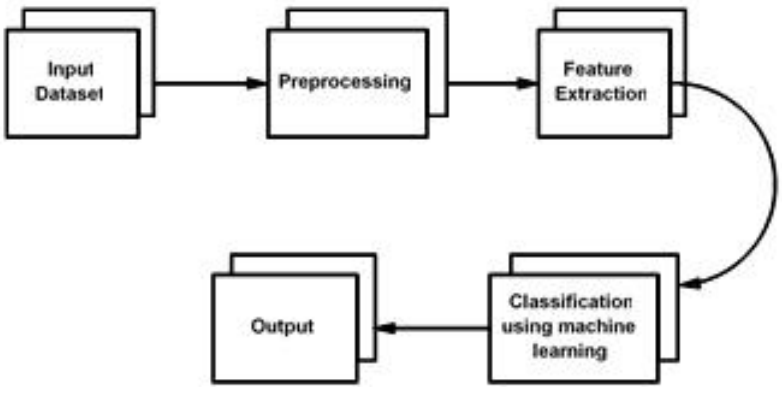

Fig. 1. Flow Diagram

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 2, February 2019

V. CONCLUSION

We reviewed several research papers on document clus-tering. Existing approaches did not work on automated text classification. There was no automatic and principled approach to feature selection. An automatic text classification approach is proposed with the aim of addressing the above-mentioned issues as well as improving accuracy and reducing the time complexity.

REFERENCES

[1] 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. 574585, 2000.
[2] J.-T. Chien, Hierarchical theme and topic modeling, IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 565578, 2016.
[3] Bernardini, C. Carpineto, and M. DAmico, Full-subtopic retrieval with keyphrase-based search results clustering, in IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intelligent Agent Technol., 2009, pp. 206213
[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. 658665.
[5] R. Xu and D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645678, 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. 148155.
[7] R. Kohavi and G. H. John, Wrappers for feature subset selection,Artif. Intell., vol. 97, no. 1, pp. 273324, 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. 574585, 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. 490499.
[10] K. Lagus and S. Kaski, Keyword selection method for characterizing text document maps, in Int. Conf. Artificial Neural Networks (ICANN), 1999, pp. 371376.