Survey on Feature Selection in Document Clustering

MS. K.Mugunthadevi
M.Phil scholar,
P.S.G.R. Krishnammal College for Women, Coimbatore, India.

MRS. S.C. Punitha
HOD, Department of Computer Science,
P.S.G.R. Krishnammal College for Women, Coimbatore, India.

Dr..M. Punithavalli
Director, Department of Computer Science,
Sri Ramakrishna college of Arts and Science for Women, Coimbatore, India.

***Abstract*-----Text mining is to research technologies to discover useful knowledge from enormous collections of documents, and to develop a system to provide knowledge and to support in decision making. Basically cluster means a group of similar data, document clustering means segregating the data into different groups of similar data. Clustering is a fundamental data analysis technique used for various applications such as biology, psychology, control and signal processing, information theory and mining technologies. Text mining is not a stand-alone task that human analysts typically engage in. The goal is to transform text composed of everyday language into a structured, database format. In this way, heterogeneous documents are summarized and presented in a uniform manner. Among others, the challenging problems of text clustering are big volume, high dimensionality and complex semantics.**

Keywords: text mining, feature selection, information retrieval, ontology, document clustering

I. INTRODUCTION

Document clustering is the task of automatically organizing text document into meaning full cluster or group, In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics. It is one of the most important tasks in text mining. There are several number of technique launched for clustering documents since there is rapid growth in the field of internet and computational technologies, the field of text mining have a abrupt growth, so that simple document clustering to more demanding task such as production of granular taxonomies, sentiment analysis, and document summarization for the scope of devolving higher quality information from text.

The problem of document clustering is generally defined as follows [31] Given a set of documents, would like to partition them into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. Documents are represented using the vector space model which treats a document as a bag of words [10]. A major characteristic of document clustering algorithms is the high dimensionality of the feature space, which imposes a big challenge to the performance of clustering algorithms. They could not work efficiently in high dimensional feature spaces due to the inherent sparseness of the data. Next challenge is that not all features are important for document clustering, some of the features may be redundant or irrelevant and some may even misguide the clustering result [32], especially there are more irrelevant features than relevant ones.

Feature selection can be a powerful tool for simplifying or speeding up computations, it can improves the text clustering efficiency and performance in ideal case, in which features are selected based on class information. Feature selection not only reduces the high dimensionality of the feature space, but also provides better data understanding, which improves the clustering result [32]. The selected feature set should contain sufficient or more reliable information about the original data set. For document clustering, this will be formulated into the problem of identifying the most informative words within a set of documents for clustering. It is widely used in supervised learning, such as text classification. It is reported that feature selection can improve the efficiency and accuracy of text classification algorithms by removing redundant and irrelevant terms from the corpus [8].

Feature selection, an effective dimensionality reduction technique, is an essential pre-processing method to remove noisy features. The entropy measure is suitable for selecting the most important subset of features because it is invariant with number of dimensions, and is affected only by the quality of clustering. Extensive performance evaluation over synthetic, benchmark, and real datasets show its effectiveness [21]. In iterative feature selection [9] method clustering are iteratively performs and feature selection in a unified framework.

Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset.

The remainder of this paper is organized as follows. Section II discusses some of the earlier proposed research work on text document clustering. Section III provides a fundamental idea on which the future research work focuses on. Section IV concludes the paper with fewer discussions.

II. RELATED WORKS

Berry Michael W et al., [33] put forward the text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value.

P. Bradley et al., [19] projected, that in current trend, extracting information from the World Wide Web has been much familiar among all. Information extraction system defined as a system that “automatically identifies predefined set of related items" [19], since a lot of Web data are found in HTML pages. Since we use HTML, the extraction process requires fetching a Web document, cleaning it up using a syntactic normalization algorithm, and then, locating “objects of interest” in this Web page. This is done by first locating the minimal object-rich sub tree. Finally, the set of objects is refined to eliminate irrelevant objects. Zamir and colleagues [35] present a system that automatically extract data from large data-intensive Web sites their “data grabber” explores a large Web site and infers a model for it, describing it as a directed graph with nodes describing classes of structurally similar pages and arcs representing links between these pages. After pinpointing classes of interest, a library of wrappers can be generated, one per class with the help of an external wrapper generator and appropriate data can be extracted.

Shen huang et al., [18] projected that web document clustering propose a novel feature co-selection, which is called multitype feature co-selection for clustering (MFCC). MFCC uses intermediate clustering results in one type of feature space to help the selection in another type of feature spaces. Feature coselection is implemented iteratively and can be well integrated into an iterative clustering algorithm. This is done by using the intermediate clustering result in one feature space as additional information to enhance the feature selection in another space. As a result, we produce better clusters in each space.

Sun Park and others [1] proposed the document clustering methods using weighted semantic features and cluster similarity is done by using NMF(non negative matrix factorization). Similarity between the clusters and document has the following advantages. First, it can easily group documents with the major topics of document using clustering based weighted semantic features [1]. Second, it can improve the quality of document clustering in view of the fact that reassigning cluster will remove deflection documents in cluster easily. This proposed method has better performance than other document clustering method using NMF (nonnegative matrix factorization).

Marcelo N. Ribeiro and others [30] gave an idea of a local feature selection approach for partitional hierarchical text clustering. Each cluster derived by the proposed method is represented by a different subset of feature, the local approach was compared to the global feature selection approach for the bisecting K-means. It was observed that the local approach obtained good precision even for few selected terms. We also performed experiments by using the ZOOM-IN method to automatically define the number of selected features in each iteration of the partitional algorithm. The results obtained by the ZOOM-IN were suitable, because it proved the need for feature selection in text clustering and showed the benefits in select features locally. Finally, the proposed method will be evaluate ranking based on entropy [34].

Thangamani.M and P.Thangaraj et al., [2] suggested that integrated clustering and feature selection scheme for text document were used group up the text documents with reference to its similarity .The feature selection is a method that eliminates the redundant and irrelevant items from the text document contents. Statistical methods used in feature selection algorithm. The integrated semantic clustering and feature selection method was proposed to improve the clustering and feature selection mechanism with semantic relation of the text documents and ontology was used to represent the terms and concept relationship [2]. Their work shows the scheme reduces the cube size and feature selection process also produced more accurate results.

Andreas Hotho et al., [14] says that text clustering and classification are two important approaches to organize textual information, e.g. from the World Wide Web. Ontology plays a central role in the Semantic Web and can be used to enhance existing technologies from machine learning and information retrieval. Ontology Learning aims at semi automatically building ontologies from a given text corpus with a limited human effort and improves the quality of the learned taxonomies by using natural language processing techniques.

Young-Woo and others [22] say ontology integration is required when attempting to integrate different source of information with differing ontologies or data schemata. Machine learning techniques are promising for (semi) automatic learning of ontologies from the Web and semantic annotation of Web documents. The areas of application of machine learning for constructing the Semantic Web: ontology learning, semantic annotation, and ontology integration. Ontology learning, refers to using machine learning techniques to (semi) automatically learn ontologies from a given dataset. Semantic annotation refers to the use of machine learning to automatically semantically annotate large corpora of data according to a given ontology.

Techniques from text learning and information retrieval can be used to build ontologies (semi) automatically [11] [12]. With the help of techniques from text learning and information retrieval fields, statistically significant terms that could serve as potential concepts in domain ontology can be presented as candidate concepts words to the domain expert constructing the ontology. In order to evaluate the learnt ontology, we investigated the usefulness of the ontology for text classification. The performance of text classification has been shown to improve in the presence of conceptually represented domain knowledge such as ontologies [13]. Therefore text classification provides a good context for evaluating the results of ontology learning.

K. Raja and others [31] stated that the proposed system is designed to identify the semantic relations using the ontology. The ontology is used represent the term and concept relationship. The synonym, meronym and hypernym relationships are represented in the ontology. The concept weights are estimated with reference to the ontology. The concept weight is used for the clustering process. Statistical methods are used in the text clustering and feature selection algorithm. The cube size is very high and accuracy is low in the term based text clustering and feature selection method.
Xu, J. Xu, B [27] put forward new feature selection method for text clustering based on expectation maximization and cluster validity is proposed. It uses supervised feature selection method on the intermediate clustering result which is generated during iterative clustering to do feature selection for text clustering; meanwhile, the Davies-Bouldin's index is used to evaluate the intermediate feature subsets indirectly. Then feature subsets are selected according to the curve of the Davies-Bouldin's index. Experiment is carried out on several popular datasets and the results show the advantages of the proposed method. The main goal is to obtain the performances better than human.

Barak Chizi et al., [28] introduced dimensionality (i.e., the number of data set attributes or groups of attributes) constitutes a serious obstacle to the efficiency of most data mining algorithms. The main reason for this is that data mining algorithms are computationally intensive. This obstacle is sometimes known as the "curse of dimensionality". The objective of Feature Selection is to identify features in the data-set as important, and discard any other feature as irrelevant and redundant information. Feature Selection reduces the dimensionality of the data. Data mining algorithms can be operated faster and more effectively by using Feature Selection.

[28] There are three main approaches for feature selection: wrapper, filter and embedded. The filter approach operates independently of the data mining method employed subsequently undesirable features are filtered out of the data before learning begins. A sub-category of filter methods that will be refer to as rankers, are methods that employ some criterion to score each feature and provide a ranking. From this ordering, several feature subsets can be chosen by manually setting. The wrapper approach uses an inducer as a black box along with a statistical re-sampling technique such as cross-validation to select the best feature subset according to some predictive measure. The embedded approach is similar to the wrapper approach in the sense that the features are specifically selected for a certain inducer, but it selects the features in the process of learning.

M. Janaki et al., [29] spot about a novel feature selection method based on Ant Colony Optimization, a swarm intelligence algorithm is
proposed. Performance of the classifier is compared to the feature selected by conventional chi-square and CHIR methods. The outcome of proposed algorithm identifies better feature set than the conventional methods.

III. FUTURE ENHANCEMENT

Some interesting research topics in feature selection of potential impact in the near future. Feature selection for ultrahigh dimensional data selecting features on data sets with millions of features and explanation-based feature selection (EBFS) feature selection via explaining training samples using concepts generalized from exiting feature and knowledge.

IV. CONCLUSION

In this survey we had projected various feature selection methods, terms, limitations, advantages and available recent innovation in feature selection. We hope, that the interested readers will have broad overview of this field and several starting point for further details. Feature selection remains and will continue to be an active field that is incessantly rejuvenating itself to answer new challenges.

REFERENCES

[1] Sun Park, Dong Un An, Choi Im Cheon, "Document Clustering Method Using Weighted Semantic Features and Cluster Similarity," digitel, pp.185-187, 2010 Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, 2010

[2] Thangamani.M and P.Thangaraj," integrated clustering and feature selection scheme fo textdocuments",J.Comput.Sci.,6:536.541,DOL:10.3844/jcssp.2010.536.541,URL:http://www.thescipub.com/abstract/10.3844/jcssp.2010.536.54

[3] M. Thangamani and p. Thangaraj , "Semantic clustering with feature selection for text documents",International J. of Engg. Research & Indu. Appls. (IJERIA),ISSN 0974-1518, Vol.3, No. II (May 2010), pp 199-210

[4] M. Thangamani and p. Thangaraj , "Survey on Text Document Clustering",(IJCSIS) International Journal of Computer Science and Information Security ,Vol. 8, No. 4, July 2010

[5] Daniela M. Witten, Robert Tibshirani."A framework for feature selection in clustering", Journal of the American Statistical Association. June 1, 2010, 105(490):713-726. doi:10.1198/jasa.2010.

[6] Mahesh T R, Suresh M B and Vinayababu.M "text mining: advancements, challenges And future directions",International Journal Of Reviews In Computing, ISSN: 2076-3328, www.ijric.org, E-ISSN: 2076-3336

[7] Wen-Hui Yang, Dao-Qing Dai, and Hong Yan, Fellow, IEEE," feature extraction and uncorreleted discriminant analysis for high dimentional data", IEEE transactions on knowledge and data engineering, vol. 20, no. 5, may 2008

[8] Yanjun Li, Congnan Luo,," text clustering with feature selection by using statistical data", IEEE Transactions on Knowledge and Data Engineering,may 2008 vol: 20 no:5

[9] Tao Liu, Shengping Liu , Zheng Chen and Wei-Ying Ma,"an evaluation on feature selection for text clustering", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[10] Luiz G. P. Almeida, Ana T. R. Vasconcelos and Marco A. G. Maia ," A Simple and Fast Term Selection Procedure for Text Clustering "Seventh International Conference on Intelligent Systems Design and Applications, 0-7695-2976-3/07 © 2007 ieee, doi :10.1109/ISDA.2007.15

[11] B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining", In Proceedings of International Semantic Web Conference (ISWC), pages 264– 278,2002.

[12] A. Hotho, S. Staab, and A. Maedche. Ontology-based text clustering. In Proceedings of the IJCAI-2001 Workshop Text Learning: Beyond Supervision, Seattle,USA, August 2001.

[13] S. Bloehdorn and A. Hotho. Text classification by boosting weak learners based on terms and concepts. In Proceedings of the Fourth IEEE International Conference on Data Mining, 2004.

[14] Andreas Hotho," Using Ontologies to Improve the Text Custering and Classification Task",Knowledge and Data Engineering Group, University of Kassel, January 14, 2005

[15] Xiang Ji,Wei Xu Shenghuo Zhu," Document Clustering with Prior Knowledge", *SIGIR'06,* August 6–11, 2006, Seattle, Washington, USA,Copyright 2006 ACM 1-59593-369-7/06/0008

[16] Magnus Rosell KTH CSC," Introduction to Information Retrieval and Text Clustering" August 1, 2006, ISBN 0- 201-39829-X.

[17] Andreas Hotho and Alexander Maedche and Steffen Staab," Ontology-based Text Document Clustering",Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany

[18] shen huang, zheng chen, yong yu, and wei-ying ma," multitype features coselection for web document clustering", ieee transactions on knowledge and data engineering, vol. 18, no. 4, april 2006, 1041-4347/06/$20.00 _ 2006 ieee published by the ieee computer society

[19] P. Bradley, U. Fayyad, and C. Reina, " Scaling clustering algorithms to large databases, " In Proc. of KDD-1998, New York, NY, USA, August 1998, pages 9–15, Menlo Park, CA, USA, 1998. AAAI Press.

[20] Dino Ienco,Rosa Meo," Exploration and Reduction of the Feature Space by Hierarchical Clustering"Dipartimento di Informatica, Universit`a di Torino,Italy,related:http://www.siam.org/proceedings/datamining/2008/dm08

[21] Manoranjan Dash ,Kiseok Choi ,Peter Scheuermann ,Huan Liu," Feature Selection for Clustering – A Filter Solution" Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)0-7695-1754-4/02 © 2002 IEEE

[22] Young-Woo Seo,Anupriya Ankolekar,Katia Sycara," Feature Selections for Extracting Semantically Rich Word for Ontology Learning" CMU-RI-TR-04-18. March 2004.

[23] Liping Jing, Lixin Zhou, Michael K, Ng, Joshua Zhexue Huang, " Ontology-based Distance Measure for Text Clustering " DASFAA'07 Proceedings of the 12th international conference on Database systems for advanced applications, ISBN: 978-3-540-71702-7

[24] Liping Jing," Survey of Text Clustering", Department of Mathematics, The University of Hong Kong, HongKong, China, , ISBN: 7695-1754-4/02

[25] Huan Liu, Senior Member, IEEE, and Lei Yu, Student Member, ieee," Toward Integrating Feature Selectio Algorithms for Classification and Clustering", ieee transactions on knowledge and data engineering, vol. 17, no. 4, april 2005

[26] Mitra, P. Murthy, C.A. Pal, S.K. ," Unsupervised feature selection using feature similarity ", Pattern Analysis and Machine Intelligence, IEEE Transactions on Mar 2002 ,Vol 24 Issue: 3

[27] Xu, J. Xu, B. Zhang, W. Cui, Z. Zhang, W."A new feature selection method for text clustering ",wuhan university journal of natural sciences, 2007, vol 12; number 5, pages 912-916
[28] Barak Chizi (Tel-Aviv University, Israel); Lior Rokach (Ben-Gurion University, Israel); Oded Maimon (Tel-Aviv University, Israel) "a survey of feature selection techniques"1888-1895 pp. John Wang (Ed.) (Montclair State University, USA), DOI:10.4018/978-1-60566-010-3.ch289, ISBN13:9781605660103,2009
[29] M. Janaki Meena,K.R. Chandran,J. Mary Brinda," integrating swarm intelligence and statistical data forfeature selection in text categorization" ©2010 International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 11
[30] Marcelo N. Ribeiro, Manoel J. Neto, Ricardo B. Prudêncio," Local Feature Selection in Text Clustering", Advances in Neuro-Information Processing, Springer-Verlag Berlin, Heidelberg ©2009 ISBN: 978-3-642-03039-0 doi>10.1007/978-3-642-03040-6_6
[31] Prof. K. Raja , C. Prakash Narayanan, "Clustering Technique with Feature Selection for Text Documents", Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010 6 February 2010, Chennai, India.
[32] Fabrizio Sebastiani "Machine Learning in Automated Text Categorization" ACM Computing Surveys, Vol. 34, No. 1, March 2002
[33] Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.
[34] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering – a filter solution. In Proceedings of the 2002 IEEE International Conference on Data Mining, pages 115–122, 2002.
[35] Zamir, O.Etzioni, "Web Document Clustering, A Feasibility Demonstration, " in Proceedings of the 21st International ACM SIGIR Conference on Research and Development.