A SURVEY ON TEXT CLASSIFICATION METHODS

[1]Arathy Chandran, [2]Aswathy M R
[1]PG Student, [2]Assistant Professor
[1]Department of Computer Science & Engineering,
[1]Vidya Academy of Science & Technology,
Thalakkottukara P. O, Thrissur, Kerala, India

Abstract— Text classification method is the task of choosing correct domain or class label for a given text document or it is extraction of relevant information from large collection of text documents. It assigns one or more classes to a document according to their content. This can be done or algorithmically and manually. There are several text classification methods and this paper provides a review of several text classification methods and focus more on ontology based text classification method.

Index Terms—Text Classification, Class, Label, Ontology.

I. INTRODUCTION

Text is the most direct source of human knowledge. Human beings write texts concerning what they see and think about the world. Thus, it is this descriptive data that facilitate humans to share and exchange their knowledge. The amount of online text data has grown drastically in recent years, due to the increase in popularity of the World Wide Web. Hence, it is no longer practical for a human being to analyze the articles or even sort it into categories. The classification of textual data, thus gains increasingly high significance. Text classification also called as text categorization is the task of assigning a document to one or more predefined categories or classes. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. If not specified, text classification is implied.

A conventional text classification model consists of text documents as the input, process with natural language processing, feature extraction, feature weighting, feature reduction, classification engine, and then being classified into relevant classes or categories. Most of the traditional text classification systems are key word based without intelligent features, and they provide erroneous result in various applications. Such methods mainly concentrate on the frequency of occurrence of a term rather its relationship or type of dependency.

Advancing the accuracy of text classification is the main intention of using ontology for the classification. The definition of Ontology is that it is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them. Ontology represent semantic relationships of the terms that can be used to correctly identify the subject of a document [12].

II. RELATED WORKS

The existing works in the area of text classification has been discussed here.

Dr. Ghayda A. Al-Talib and Hind S. Hassan [3] presented a paper on SMS classification using Tf-idf weighting. The paper presents a Tf-idf weighting model. The method is based on statistical evaluation of significance of a word for an SMS categorization problem. The said method will classify mobile SMS into predefined categories. Initially, all sms are transformed to text documents. Next, documents are processed to select the noteworthy words that carry the meaning and neglect the words that are unimportant. Processing stage includes steps such as replacement of abbreviations, stop words removal, part of speech tagging and the stemming technique. These processes are discussed in detail. After preprocessing stage vector space model is prepared and weight is assigned to each term. The paper shows that it improves the accuracy of the classification method significantly.

Zhang Yun-tao *et al.* [4] proposed an improved TF-IDF approach for text classification. The conventional TF-IDF approach is used to weigh every word in the document according to how distinctive it is. Otherwise TF-IDF method captures the relevancy among words, text documents and particular categories. The system uses confidence, support and characteristic words to improve the recall and precision of text classification. Confidence is the measure of certainty to establish a particular class by a particular feature word. The potential usefulness of particular feature word is represented by support. Synonyms defined by a word list are processed in this enhanced Tf-idf approach. The results show that the improved tf-idf approach improves the accuracy of text classification compared with the conventional approach of tf-idf.

In paper [5], Akiko Aizawa presents the information theoretic interpretation of tf-idf measures. It gives the mathematical explanation of the "probability-weighted amount of information (PWI)" which is a measure of specificity of the terms in the documents. The specificity is based on an information theoretic view of retrieval events. A measure of specificity is defined based on the amount of information or the entropy of terms. This measure adds up the deviations from randomness of the occurrences of terms. The PWI is expressed as a product of the occurrence term probabilities and their quantities of information, and corresponds well with the conventional tf-idf measures. By evaluating the PWI values of a series of query terms under various probability suppositions, they have shown that the vector-space-oriented view of the original tf–idf can fruitfully be linked to probability-oriented views.

Agarwal C.C *et al.* [11] presented a paper about usage of partial supervision for the text categorization. Paper points out the pros of building text categorization systems by using supervised clustering techniques. It also discusses about the cons of completely unsupervised techniques, that it has complication in separating sufficiently fine-grained classes of documents relating to a coherent subject matter. In order to supervise the process of creation of a set of co-related clusters by using a seed-based technique, they have taken information from a preexisting taxonomy. They built a classification system using Yahoo! Taxonomy and have used them as base in order to create supervised clusters. Also showed that the supervised clustering method suits well to automated categorization with higher overall quality of classification.

In paper [12] Wijewickrema *et al.* discusses about the impact of using ontologies for improving the accuracy of text classification. The paper proposes a solution to reduce the number of misclassifications because of the vocabulary ambiguities of the used language, which is based on ontology. As ontology represents the relationships among the concepts and descriptions of those concepts, it is easy to find and clearly understand all possible meanings of an ambiguous term. Hence, it can be used to choose the most suitable candidate out of number of other classification results. The choice of the suitable class is done by a newly developed automatic text classification system. The proper

integration of the ontology with the automatic text classifier furnishes the goal of obtaining an ontology based automatic text classification system.

Amal Zouaq and Roger Nkambou [13] presents a semiautomatic framework that aims to construct domain concept maps from text and then to develop domain ontologies from these concept maps. Paper details the steps to transform textual resources, into concept maps pertaining to certain domain and explains how abstract structure is changed into formal domain ontology. To validate the resulting ontology, this paper also portrays an evaluation methodology and presents the results obtained in an evaluation with a corpus of text documents.

A Paper describes a work [14] which implements Text mining methods to automatically organize research papers according to their research areas. It presents a pattern based ontology text mining approach, in which the initial step is that one put paper and year of submission of the concerned paper. Then make a pattern and cluster the research articles based on the similarities in their research areas. The planned approach builds the research ontology and then applies Decision Tree Algorithm to organize the data into the disciplines using research ontology and then the resultant of classification is used to construct clusters of related data.

Liu *et al.* in [15] introduces a new methodology to extract the concepts of ontology from multiple text documents of same kind. Proposed method uses mutual information and document frequency. As an initial step perform text processing on the corpus, and then based on N-gram algorithm generates a set of candidate noun phrases and finally uses the statistical and linguistic rules to screen for the concept of ontology from candidate phrases. This method of extracting the concept of domain ontology is composed of two phrases and can extend this method for extracting the ontology concepts composed of three word or four word phrases.

Wijewickrema *et al* [16] talks about an automated system that can entirely categorize a given text document by minimizing the ambiguities in the vocabulary. The paper points out the method to further improve a tf-idf based semi-automatic (hybrid) text classification system. The basic tf-idf function used in the semiautomatic framework has been enhanced here. In order to diminish the vocabulary ambiguities, a domain-ontology has been used. Proposed framework has several stages of classification and the obtained results are analyzed using binary logistic regression. In [17] Rozeva, points out an automatic text classification method by implementing domain ontologies. The idea consists of extracting the basic topics for training the classifier via unsupervised machine learning on a text corpus and additional arrangement of the document vectors to concepts of the ontology. The results got by classification of the text documents supervised by e-governance ontology with numerous machine learning algorithms showed adequate match of their content to the ontology concepts.

Suha S. Oleiwi, and Azman Yasin in [18], proposed a method for classifying scientific papers to multiple categories using ontologies. They points out that one of the most important problems associated with classification procedure is the high dimensionality which caused by the number of training set, the training set effect on the arrangement result. The core objective of the work is to decrease the training set by using ontology as a classifier and a model is proposed to organize the scientific papers into a proper set of the clear categories, by using the semantic relation between concepts on ontology with respect to the location of the concept in the respective document. Similarity is calculated to decide if the concept is significant for any particular document or not. Proposed model works only on the ontology concepts relation. So there is a need for a professional when the ontology model is constructed.

In another paper, Nazia Ilyas Baig *et al* [20], talks about Word Sense Disambiguation tool for ontology based text classification. When the training set is not available to train the classifier; Ontology provides us with knowledge that can be powerfully used for classification without using training sets. For categorizing the documents using ontology , there is a need to distinguish the class or the concepts to classify the document. In order to capture the relationship between words WordNet is used in the proposed method. As it is seen that WordNet alone is not sufficient to remove Word Sense Disambiguation (WSD), in the proposed method Lesk algorithm is used to deal with WSD. The proposed method is leveraging the strengths of ontology, WordNet and Lesk Algorithm for improving the text document classification.

III. CONCLUSION

From the literature survey done so far, it was observed that the text mining studies are gaining more importance recently due to the availability of the increasing number of the electronic documents from a large variety of resources. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification and summarization. Text classification is the task of automatically sorting a set of documents into categories from a predefined set. In this paper, it has been reviewed several text classification papers. This review would help future research based on the past studies. It was verified from the study most of the traditional text classification systems are keyword based without intelligent features, providing inaccurate results in various applications. They mainly concentrate on Tf-idf methodology. It considers only the occurrence of a term rather than its relationship or type of dependency with its respective document. Such methodologies some time generate inaccurate results. Ontology is the best solution for this problem. It takes positively dependent terms of a document for the classification purpose. It is therefore more useful to build intelligent application because computer application is usually developed for particular target domains. A well-constructed ontology can help develop a knowledge-based information search and management system more effectively, such as Web search engine, semiautomatic text classification, and content management system.

IV. ACKNOWLEDGMENT

We wish to record our indebtedness and thankfulness to all those who helped us to prepare this paper and present it in a satisfactory way. Our sincere thanks to Dr. Sudha Balagopalan, our Principal, for providing us all the necessary facilities. We are also thankful to Ms. Sunitha C, Head of Department of Computer Science and Engineering, for encouragement. Last but not the least, we wish to thank our family and friends for supporting and encouraging us throughout the work of this paper.

REFERENCES

[1] Edward H.Y. Lim, James N.K. Liu, and Raymond S.T. Lee, Knowledge Seeker Ontology Modeling for Information Search and Management Intelligent Systems, Reference Library volume 8.

[2] James N. K. Liu, Yu-Lin He, Edward H. Y. Lim, Xi-Zhao Wang, "A New Method for Knowledge and Information Management: Domain Ontology Graph Model", IEEE Transactions On Systems, Man, And Cybernetics: Systems, Vol. 43, No. 1, January 2013, pp.115 - 127.

[3] Al-Talib, Ghayda A., and Hind S. Hassan, "A Study on Analysis of SMS Classification Using TF-IDF Weighting", International Journal of Computer Networks and Communications Security (2013), Vol.1, No.5, pp.189 - 194.

[4] Yun-tao Z, Ling G, Yong-cheng W, "An improved TF-IDF approach for text classification", Journal of Zhejiang University Science A (2005) Vol.6, No.1, pp.49-55.

[5] Aizawa and Akiko, "An information-theoretic perspective of tf–idf measures", Information Processing & Management (2003), Vol.39, No.1, pp.45-65.

[6] Wu, Wentao, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. "Probase: A probabilistic taxonomy for text understanding", In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp.481-492. ACM, 2012.

[7] Jindal, Rajni, Ruchika Malhotra, and Abha Jain, "Techniques for text classification: Literature review and current trends", Webology (2015), Vol.12, No.2, pp 1.

[8] Korde, Vandana, and C. Namrata Mahender, "Text classification and classifiers: A survey", International Journal of Artificial Intelligence & Applications (2012), Vol.3, No. 2 pp.85.

[9] Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T, "Text classification from labeled and unlabeled documents using EM", Machine learning, Vol.39, No.(2-3), pp.103 - 134.

[10] Faraz, Ahmed, "An Elaboration Of Text Categorization And Automatic Text Classification Through Mathematical And GraphicalModelling", An International Journal (CSEIJ), Vol.5, No.2/3, June 2015.

[11] Aggarwal, C.C., Gates, S.C. and Yu, P.S., 2004, "On using partial supervision for text categorization", IEEE Transactions on Knowledge and data Engineering, Vol.16 No.2, pp.245 - 255.

[12] Wijewickrema, Chaaminda Manjula, "Impact of ontology for automatic text classification" (2014).

[13] Amal Zouaq and Roger Nkambou 2008, "Building domain ontologies from text for educational purposes", IEEE Transactions on learning technologies, Vol.1, No.1, pp.49-62.

[14] Pandey, Jay Prakash, Shrikant Lade, and Manish Kumar Suman, "Automatic Ontology Creation For Research Paper Classification.", Int. J. Engg. Res. & Sci. & Tech (2013).

[15] Liu, Yuefeng, Minyong Shi, and Chunfang Li, "Domain ontology concept extraction method based on text", In Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on, pp. 1-5. IEEE, 2016.

[16] Wijewickrema, Chaaminda Manjula, and Ruwan Gamage, "An ontology based fully automatic document classification System using an existing semi-automatic system." (2013).

[17] Rozeva.A, "Classification of text documents supervised by domain ontologies", ATI-Applied Technologies & Innovations, 8(3), pp.1-12.

[18] Suha S. Oleiwi, and Azman Yasin, "Classify the Scientific Paper to Multi Categories Using Ontology", International Conference on IT and Intelligent Systems (ICITIS'2013) August 28-29, 2013 Penang (Malaysia).

[19] Vogrinčič S, Bosnić Z, "Ontology-based multi-label classification of economic articles". Computer Science and Information Systems. 2011, Vol.8, No.1, pp.101-19.

[20] Nazia Ilyas Baig, Gresha Bhatia, "WSD Tool for Ontology-based Text Document Classification", IJAIS Proceedings on International Conference and workshop on Advanced Computing ICWAC (3):1-6, June 2013.