A Survey on Document Categorization based on Keyword and Key Phrase Extraction using Various Algorithms

Prerna Madaan[1] Gurmeet Singh[2]
[1]M. Tech Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Kurukshetra institute of Technology& Management

Abstract— Text classification is the process of classifying the text documents based on words, phrases and word combination with respect to set of predefined categories. Text classification has many applications such as mail routing, email filtering, news classification etc. and the various institutions and industries are converting their documents into electronic text files. Keyword Nets using TF-IDFs .The words which have highest similarity or frequency are taken as keywords. In this survey paper we described various algorithms for document categorization.

Key words: Text Mining, Naïve Bayes, Support Vector Machines (SVM), TF-IDF, k-Nearest Neighbour (k-NN), Vector Space Model, Weight Adjusted k-Nearest Neighbour (WAKNN)

I. INTRODUCTION

Document classification/categorization is a problem in information science. The task is to assign an electronic document to one or more categories, based on its contents. Document categorization can be done into two ways: supervised document categorization where some external mechanism (such as human feedback) provides information on the correct categorization for documents, and unsupervised document categorization, where the categorization must be done entirely without interference of external information.

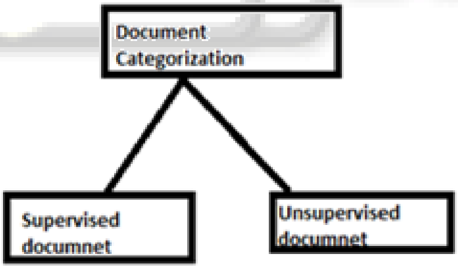

Fig. 1: Document Categorization

Document categorization [1] [2] refers to the process of deriving high-quality information from text. 'High quality' in Document categorization means that information extracted should be beneficial for the user, and according to their interest. Document categorization [3] is similar to data mining, except that data mining tools [4] are designed to handle structured data from databases, but Document categorization can also work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. Document categorization also known as Intelligent Document Analysis, Document Data Mining. By categorization that text required knowledge can be extracted which will be very useful. Knowledge from text usually refers to some combination of relevance, novelty, and interestingness. Document analysis involves information extraction, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, data mining techniques including link and association analysis. A typical application of Document categorization is to scan given set of documents written in a natural language and either to model them for predictive categorization. There is a need to construct automatic text classifier using pre-classified sample documents whose accuracy and time efficiency is much better than manual text classification. In this paper we summarize Document categorization techniques that are used to classify the text documents into predefined classes [5].

II. DOCUMENT CATEGORIZATION OR TEXT CLASSIFICATION PROCESS

The stages of Document Categorization are discussing as following points.
Fig. b Text Classification Process

A. *Documents Collection:*

In this first step of process firstly we are collecting the different types of format of document like Ms Word, pdf. Doc, web content etc.

B. *Pre-Processing:*

The first step of pre-processing is used to present the text documents into clear word format. The documents prepared for next step in document categorization are represented by a great amount of characteristics. Commonly the following steps are:

1) Tokenization: A document is treated as a string, and then divided into a list of tokens.
2) Removing stop words: Stop words such as “to”, the “or”, etc. are mostly occurring, so these unnecessarily words need to be removed.
3) Stemming word: Applying the stemming algorithm that converts different form of words into similar word form. This step is very essential for use. Like: connection to connect, computing to compute.

C. *Indexing:*

The documents representation is one of the most relevant pre-processing techniques that are used to overcome the complexity of the documents and make them easier to use. Firstly all the documents have to be converting from the full text version to a document vector. The most commonly used document representation is called Vector Space Model (VSM). In which all the documents are showed by vectors of words. Usually, one has a collection of all documents and that is showed word by word document Matrix.

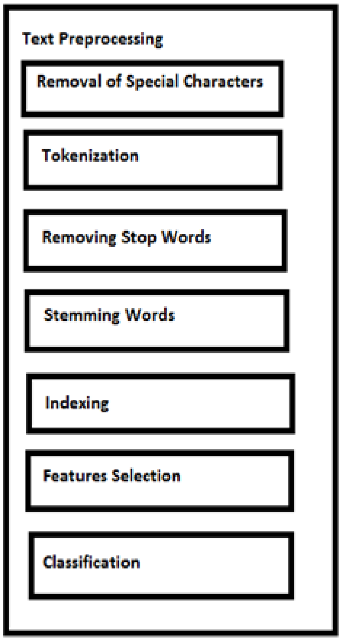

Fig. 2: text Classification Process

D. *Feature Selection:*

After pre-processing and indexing the important step of text classification, is feature selection [2] to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space. Many feature evaluation metrics have been notable among which are information gains (IG), term frequency, Chi-square. But FS of association word mining is more efficient than IG and document frequency. .Other various methods are presented like [58] sampling method which is randomly samples roughly features and then make matrix for classification. By considering problem of high dimensional problem [59] is presented new FS witch use the genetic algorithm (GA) optimization.

E. *Classification:*

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, K-nearest neighbour(KNN), Support Vector Machines (SVMs), TF-IDF .

III. Document classification techniques

A. *K-Nearest Neighbours:*

The basic idea is to determine the category of a given query based not only on the document that is nearest to it in the document space, but on the categories of the k documents that are nearest to it. Having this in mind, the vector method can be viewed as an instance on the KNN method, where k=1. This work uses a vector based, distance-weighted matching function, as did yang, by calculating document's similarity like Vector method.

K NN classifier is a case-based learning [9] algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measure's. This method is try for many application [10] Because of its effectiveness, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k .The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good k can be selected by various heuristic techniques .to overcome this drawback [11] modify traditional KNN with different K-values for different classes rather than fixed value for all classes

Fang Lu have been try to improve performance of KNN by using WKNN [12].

A major drawback of the similarity measure used in k-NN is that it uses all features in computing distances. In many document data sets, only smaller number of the total vocabulary may be useful in categorizing documents. A possible approach to overcome this problem is to learn weights for different features (or words in document data etc.) [12] propose the Weight Adjusted k-Nearest Neighbour (WAKNN) classification algorithm that is based on the k-NN classification paradigm.

B. *Naïve Bayes*

Naïve bias method is kind of module classifier [15] under known priori probability and class conditional probability .Its basic idea is to calculate the probability that document D is belongs to class C. There are two event model are present for naive Bias [16] [17] [18] as multivariate Bernoulli and multinomial model. Out of these model multinomial model is more suitable when database is large, but there are identifies two serious problem with multinomial model first it is rough parameter estimated and problem it lies in handling rare categories that contain only few training documents. They [19] propose Poisson model for NB text classification and also give weight enhancing method to improve the performance of rare categories. Modified NB is propose [20] to improve performance of text classification, also [21] provides ways to improve naïve Bayes classification by searching the dependencies among attribute. Naïve Bayes is easy for implementation and computation. So it is use for pre-processing [22] i.e. for vectorization. Performance of naïve bias is very poor when features are highly correlated and, highly it is sensitive to feature selection so the [23] propose two metrics for NB which applied on multiclass text document.

C. *Decision Tree:*

When decision tree is used for text classification it consist tree internal node are label by term, branches departing from them are labeled by test on the weight, and leaf node are represent corresponding class labels .Tree can classify the document by running through the query structure from root to until it reaches a certain leaf, which represents the goal for the classification of the document. Most of training data will not fit in memory decision tree construction it becomes inefficient due to swapping of training tuples. To handle this issue [24] presents method which can handle numeric and categorical data. New method is proposing [25] as FDT to handle the multi-label document witch reduce cost of induction, and [26] presented decision-tree-based symbolic rule induction system for text categorization which also improves text classification. The decision tree classification method is outstanding from other decision support [27] tools with several advantages like its simplicity in understanding and interpreting, even for non-expert users. So for that it is used in some application [28]

D. Support Vector Machine:

The application of Support vector machine (SVM) method to Text Classification has been propose by [32]. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. SVM classifier method is outstanding from other with its effectiveness [5] to improve performance of text classification [20] .SVM is more capable [8] to solve the multi-label class classification.

E. Term Frequency/Inverse Document Frequency (TF-IDF):

This paper presents a new improved Term frequency/Inverse document frequency (TF-IDF) approach which uses confidence, support and characteristic words to enhance the recall and precision of text classification [16]. Synonyms defined by a lexicon are processed in the improved TF-IDF approach. It discusses and analyse the relationship among confidence, recall and precision. The experiments based on science and technology gave promising results that the new TF-IDF approach improves the precision and recall of text classification compared with the conventional TF-IDF approach. In text classification, a text document may partially match many categories. It needs to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories [31].

Classification techniques have been applied to

- Spam filtering, a process which tries to discern E-mail spam messages from legitimate emails.
- Email routing, sending an email sent to a general address to a specific address or mailbox depending on topic.
- Language identification, automatically determining the language of a text.
- Genre classification, automatically determining the genre of a text.
- Readability assessment, automatically determining the degree of readability of a text, either to find suitable materials for different age groups or reader types or as part of a larger text simplification system.

IV. COMPARATIVE OBSERVATIONS

The performance of a classification algorithm is greatly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also reduce the Quality of the result in some cases [3]. Each algorithm has its own advantages and disadvantages as described in Table.1 with their time complexity by taking considering summary from [49][52] .The works in [5] [54] compare the most common method in most cases support machine and K-nearest neighbour have better effect neural network is after then and then naïve bays is last and its evaluation index is again break –even point.

V. PROPOSED WORK

Document Classification in the proposed system can be done by using the combination of Naïve Bayes, k-NN and Support Vector Machine algorithms along with keyword dataset and training dataset which is extracted based TF-IDF values of words. The various algorithms are applied for various kinds of documents to improve the classification accuracy. They split the whole work into various modules and tasks, to improve the accuracy of the classification. Here the extracted keywords and key phrases are considered as training set data for future classification.

VI. CONCLUSION

The growing use of the textual data which needs text mining, natural language processing and machine learning techniques and methodologies to organize and extract pattern and knowledge from the documents. This survey focused on the existing literature and explored the document representation and an analysis of feature selection method and classification algorithm were presented.it was verified from the study that information Gain and Chi square are the most commonly used and well performed methods for feature selection. The existing classification methods are compared and contrasted based on various parameters namely criteria used for classification. Different algorithms perform differently depend upon the data collection.to the certain extent SVM with term weighted VSM representation schemes performs well in many text classification tasks. In this Paper almost all the techniques have been extended to the case of text data. In recent years, the advancement of web and social network technologies have leads to a tremendous interest in the classification of text documents containing links or other meta-information can significantly improve the quality of the underlying results.

ACKNOWLEDGEMENT

During the time of writing of this paper I received help from many people. In particular, I am thankful to my supervisor, Mr Gurmeet Singh at Computer Science department of Kurukshetra Institute of Technology & Management, who was very kind with his time and knowledge and guide me in each step to complete the paper. I am thankful to all my near and dear for their encouragement.

REFERENCES

[1] F. Sebastiani, “Text categorization”, Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005.

[2] A. Dasgupta, P. Drineas, B. Harb, “Feature Selection Methods for Text Classification”, KDD’07, ACM, 2007.

[3] A. Khan, B. Baharudin, L. H. Lee, K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification”, Journal of Advances Information Technology, vol. 1, 2010.

[4] F. Sebastiani, “Machine Learning in Automated Text Categorization”, ACM 2002.

[5] Y. Y. X. Liu, “A re-examination of Text categorization Methods” IGIR-99, 1999.

[6] Hein Ragas Cornelis H.A. Koster, “Four text classification algorithms compared on a Dutch corpus” SIGIR 1998: 369-370 1998.

[7] Susan Dumais John Platt David Heckerman, “Inductive Learning Algorithms and Representations for Text Categorization”, Published by ACM, 1998.

[8] Michael Pazzani Daniel Billsus “Learning and Revising User Profiles: The Identification of Interesting Web Sites”, Machine Learning, 313–331 1997

[9] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, “KNN Model-Based Approach in Classification”, Proc. ODBASE pp- 986 – 996, 2003

[10] Eiji Aramaki and Kengo Miyo, “Patient status classification by using rule based sentence extraction and bm25-knn based classifier”, Proc. of i2b2 AMIA workshop, 2006.

[11] Muhammed Miah, “Improved k-NN Algorithm for Text Classification”, Department of Computer Science and Engineering University of Texas at Arlington, TX, USA.

[12] Fang Lu Qingyuan Bai, “A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization”, IEEE 2010.

[13] Kwangcheol Shin, Ajith Abraham, and Sang Yong Han, “Improving kNN Text Categorization by Removing Outliers from Training Set”, Springer-Verlag Berlin Heidelberg 2006.

[14] Methods Ali Danesh Behzad Moshiri “Improve text classification accuracy based on classifier fusionmethods”. 10th International Conference on Information Fusion, 1-6 2007.

[15] SHI Yong-feng, ZHAO, “Comparison of text categorization algorithm”, Wuhan university Journal ofnatural sciences. 2004.

[16] D. Lewis, “Naive Bayes at Forty: The Independence Assumption in Information Retrieval”, Proc. ECML-98, 10th European Conf. Machine 1998.

[17] Vidhya. K.A G.Aghila, “A Survey of Naïve Bayes Machine Learning approach in Text Document Classification”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[18] McCallum, A. and Nigam K., "A Comparison of Event Models for Naive Bayes Text Classification".AAAI/ ICML -98 Workshop on Learning for Text Categorization

[19] Sang- Bum Kim, et al, “Some Effective Techniques for Naive Bayes Text Classification “IEEE Transactions on Knowledge and Data Engineering, Vol. 18, November 2006.

[20] Yirong Shen and Jing Jiang” Improving the Performance of Naive Bayes for Text Classification”CS224N Spring 2003

[21] Michael J. Pazzani “Searching for dependencies in Bayesian classifiers” Proceedings of the Fifth Int. workshop on AI and, Statistics. Pearl, 1988

[22] Dino Isa “Text Document Pre-Processing Using the Bayes Formula for Classification Based on the 94

[23] Bayes Jingnian Chen a, b, Houkuan Huang a, Shengfeng Tian a, Youli Qua a “Feature selection for text classification with Naïve”, China Expert Systems with Applications 36 5432–54352009

[24] Mnish Mehta, Rakesh agrwal” SLIQ: A Fast Scalable Classifier for Data Mining” 1996.

[25] Peerapon Vateekul and Miroslav Kubat, “Fast Induction of Multiple Decision Trees in Text Categorization From Large Scale,Imbalanced, and Multi-label Data”, IEEE International Conference on Data MiningWorkshops 2009

[26] D. E. Johnson F. J. Oles T. Zhang T. Goetz, “A decision-tree-based symbolic rule induction system for text Categorization”, by IBM SYSTEMS JOURNAL, VOL 41, NO 3, 2002

[27] [27]David D. Lewis and Marc Ringuette, “A comparison of two learning algorithms for text categorization”, Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US 1994.

[28] HAO CHEN, YAN ZHAN, YAN LI, “The Application Of Decision Tree In Chinese Email Classification”, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010

[29] C.Apte, F. Damerau, and S.M. Weiss “Automated Learning of Decision Rules for Text Categorization”, ACM Transactions on Information Systems, 1994

[30] Sholom M. Weiss Nitin Indurkhya, “Rule-based Machine Learning Methods for Functional Prediction”, Journal of Artificial Intelligence Research 3 383-403 1995

[31] Chih-Hung Wu “Behavior-based spam detection using a hybrid method of rule-based Techniques and neural networks”, Expert Systems with Applications 36 4321– 4330 2009

[32] Joachims, T. “Text categorization with support vector machines: learning with many relevant features”. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137–142 1998.

[33] Loubes, J. M. and van de Geer, S “Support vector machines and the Bayes rule in classification”, Data mining knowledge and discovery 6 259-275.2002

[34] Chen donghui Liu zhijing, “A new text categorization method based on HMM and SVM”, IEEE2010

[35] Yu-ping Qin Xiu-kun Wang, “Study on Multi-label Text Classification Based on SVM” Sixth International Conference on Fuzzy Systems and Knowledge Discovery 2009

[36] Dagan, I., Karov, Y., and Roth, D. “Mistake-Driven Learning in Text Categorization.” In Proceedings of CoRR. 1997

[37] MIgual E .Ruiz, Padmini Srinivasn, “Automatic Text Categorization Using Neural networks”, Advaces in Classification Research, Volume VIII.

[38] Cheng Hua Li , Soon Choel Park “An efficient document classification model using an improved back propagation neural network and singular value decomposition”, Expert Systems with Applications, 3208–3215, 2009

[39] Hwee TOU Ng Wei Boon Goh Kok Leong Low, “Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization”, SIGIR 97 Philadelphia PA,

[40] Amy J.C. Trappey a, Fu-Chiang Hsu a, Charles V. Trappey b, Chia-I. Lin “Development of a patent document classification and search platform using a back-propagation network”, Expert Systems with Applications 31 755–765 2006

[41] Yiming Yang And Christopher G. Chute Mayo Cllnic “An Example-Based Mapping Method For Text Categorization And Retrieval” ACM Transactions On Information Systems, Vol. 12, No 3, Pages 252-277, July 1994

[42] Yiming Yang Christopher G. Chute “A Linear Least Squares Fit Mapping Method For Information Retrieval From Natural Language Texts” Acres De Coling-92 Nantes, 23-28 AOUT 1992

[43] Li, Y. H. and Jain, A. K. “Classification of text documents”. The Computer Journal, 537–546. 1998.

[44] Larkey, L. S. and Croft, W. B. “Combining classifiers in text categorization”. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (Zurich, CH, 1996), pp. 289–297 1996

[45] O. Zaiane, and M. Antonie, “Text Document Categorization by Term Associaton”, Proceedings of ICDM 2002, IEEE, , pp.19-26 2002

[46] Supaporn Buddeewong1 and Worapoj Kreesuradej” A New Association Rule-Based Text Classifier Algorithm”, Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, 2005

[47] S. M. Kamruzzaman, Chowdhury Mofizur Rahman: “Text Categorization using Association Rule and Naive Bayes Classifier” CoRR, 2010

[48] Mohammad Masud Hasan and Chowdhury Mofizur Rahman,” Text Categorization Using Association Rule Based Decision Tree”, Proceeding of the 6th International Conference on Computer and Information Technology (ICCIT), pp 453-456, Bangladesh, 2003

[49] Sholom M. Weiss, Chidanand Apte, Fred J. Damerau, David E. Johnson, Frank J. Oles, Thilo Goetz, and Thomas Hampp, IBM T.J. Watson Research Center “Maximizing Text-Mining Performance” 1094-7167/99 IEEE INTELLIGENT SYSTEMS. 1999

[50] Songbo Tan ”An improved centroid classifier for text categorization” Expert Systems with Applications xxx 2007

[51] Eui-Hong (Sam) Han and George Karypis “Centroid-Based Document Classification: Analysis & Experimental Results” PKDD '00 Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery Springer-Verlag London, UK ©2000.

[52] B S Harish, D S Guru, S Manjunath ” Representation and Classification of Text Documents: A Brief Review” IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition”RTIPPR, 2010.

[53] Shi Yong-Feng, Zhao Yan-Ping in Wuhan “ Comparison of Text Categorization Algorithms ” University Journal of Natural Sciences 2004

[54] Yiming Yang “An Evolution of statistical Approaches to Text Categorization” Information Retrieval 1, 69-90 1999.

[55] Kjersti Aas and Line Eikvil “Text Categorization: A Survey” Report No. 941. ISBN 82-539-0425-8. , June, 1999.

[56] B S Harish, D S Guru, S Manjunath “Representation and Classification of Text Documents: A Brief Review” IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, 2010.

[57] Su-Jeong Ko and Jung-Hyun Lee “Feature Selection Using Association Word Mining for Classification “H.C. Mayr et al. (Eds.): DEXA 2001, LNCS 2113, pp. 211–220, 2001.

[58] Anirban Dasgupta “Feature Selection Methods for Text Classification “KDD’07, August 12–15, 2007.

[59] Wei Zhao “A New Feature Selection Algorithm in Text Categorization “International Symposium on Computer, Communication, Control and Automation 2010.