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

Yoganand. C. S[1],
[1]M.E (CSE) Student, Adithya Institute of Technology, Coimbatore

Praveen. N[2],
[1]M.E (CSE) Student, Adithya Institute of Technology, Coimbatore

Saranya. N[3],
[1]M.E (EEE) Student, Vellalar College of Engineering and Technology, Erode

Ganesh Karthikeyan. V[4]
Assistant Professor, Adithya Institute of Technology, Coimbatore

***Abstract*-The various institutions and industries are converting their documents into electronic text files. These documents plays a vital role in every part of our life. The documents may contains applications, personal documents, properties documents etc. The categorization of the text documents are really makes a very big issue. In this paper we propose the various techniques for the document classification process. These documents may be in the form of supervised, unsupervised or semi-supervised documents. The supervised documents are the standard documents which are contains the proper format of data. They can be classified by using the Naïve Bayes model with the help Hidden Markov Model (HMM). The major classification of these documents can be done by using the extraction of keyword and key phrase from the base documents. The extracted keyword and key Phrases are used as a training set for the further document classification along with the training dataset. The keyword extraction can be done based on the Word count method and Porter stemming algorithms. Further documents can be classified using Naïve Bayes and Support Vector Machine (SVM) methods with k-Nearest Neighbour (k-NN) clustering method.**

Keywords-Support Vector Machines (SVM), Hidden Markov Model (HMM), k-Nearest Neighbour (k-NN), Text categorization, mapping models.

I. INTRODUCTION

All institutions and private companies nowadays keep their files in electronic format in order to reduce the paperwork and, at the same time, provide instant access to the information contained. Document clustering and classification in one of the most important text mining methods that are developed to help users effectively navigate, summarize and organize text documents [1]. Document classification can be defined as the task of automatically categorizing collections of electronic documents into their annotated classes based on their contents. Recent years, this has become important due to the advent of large amounts of data in digital form. Document classification in the form of text classification systems have been widely implemented in numerous applications such as spam filtering, emails categorizing, directory maintenance and ontology mapping.

An increasing number of supervised classification approaches have been developed for various types of classification tasks, such as rule induction (Apte, Damerau, &Weiss, 1994; Provost, 1999), k-nearest neighbor classification (Han, Karypis, & Kumar, 1999), maximum entropy (Nigam, Lafferty, & McCallum, 1999), artificial neural network (Diligenti, Maggini, & Rigutini, 2003a, 2003b), support vector machines (Isa, Lee, Kallimani, & Rajkumar, 2008a, 2008b; Joachims, 1998; Lin, 1999), and Bayesian classification (Domingos & Pazzani, 1997; Eyheramendy, Genkin, Ju, Lewis, & Madigan, 2003; Kim, Rim, Yook, & Lim, 2002; McCallum & Nigam, 2003; O'Brien & Vogel, 2003; Provost, 1999; Rish, 2001). Besides the supervised classification approaches, the unsupervised clustering approaches, such as self-organizing map (Adami, Avesani, & Sona, 2005; Hartley, Isa, Kallimani, & Lee, 2006; Isa, Kallimani, & Lee, 2009; Wang, 2001) have also been widely implemented in segmenting data into groups for further analysis and processing.

Data mining is useful in discovering implicit, potentially valuable information or knowledge and previously unknown from large datasets. Text Document classification denotes the test of assigning raw text documents to one or more pre-defined categories. This is a direct concept from machine learning, which denotes the declaration of a set of labelled categories as a way to represent the documents, and a statistical classifier trained with a labelled training set [2]. Among these approaches, Bayesian classification has been widely implemented in many real world applications due to its relatively simple training and clustering algorithms.

One of the outstanding features of Bayesian classification as compared to other classification approaches is its ability and simplicity in handling raw text data directly, without requiring any pre-process to transform text data into a representation suitable format, typically in

numerical form, as required by most of the successful and highly accurate text classification approaches, such as by the use of k-nearest neighbor (k-NN) and support vector machines (SVM) classifiers. As a trade-off to its simplicity, Bayesian classification has been reported as one of the poorest-performing classification approaches by many research groups through extensive experiments and evaluations (Brücher, Knolmayer, & Mittermayer, 2002; Yang & Liu, 1999).

Each of the document classification schemes previously mentioned has its own unique properties and associated problems. The decision tree induction algorithms and the rule induction algorithm are simple to understand and interpret. However, these algorithms do not work well when the number of distinguishing features between documents is large. The k-NN algorithm is easy to implement and shows its effectiveness in a variety of problem domains.

The words that contained in text documents which match any word from the list of stop words will not be taken into account for both the training and classifying processes [4]. There is a potential drawback of stop word elimination, where certain words which are considered as stop words for a particular dataset (domain), but can be highly informative features for another dataset (domain) (Takamura, 2003). Besides the simple stop word elimination technique, there are several statistical methods for feature selection which have been introduced as pre-processes for Bayesian text classification. These methods provide a measure for usefulness of each individual word in the classification task.

II. DOCUMENT CLASSIFICATION TECHNIQUES

The document classification tasks can be divided into three parts: unsupervised document classification (also known as document clustering), supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, where the classification must be done entirely without reference to external information, and semi-supervised document classification, where parts of the documents are labelled by the external mechanism.

A. Bayesian classification approach

The conventional Bayesian classification approach performs its classification tasks starting with the initial step of analysing text document by extracting words which are contained in the document to generate a list of words (Isa, Lee, & Kallimani, 2008). The list of words is constructed with the assumption that input document consists of words w1, w2, w3 . . . . …wn-1, wn, where the length of the document (in terms of number of words) is n.

Based on the list of words, the trained Bayesian classifier calculates the posterior probability of a particular word of the document being annotated to a particular category by using the formula which is shown in Eq. (1), since each word in the input document contributes to the document's categorical probability [17].

$$\Pr(Category|Word) = \frac{\Pr(Word|Category)\cdot\Pr(Category)}{\Pr(Word)} \quad (1)$$

The derived equation above shows that by observing the value of a particular word, wj, the prior probability of a particular category, Ci, Pr(Ci) can be converted to the posterior probability, Pr(Cijwj), which represents the probability of a particular word, wj being a particular category, Ci. The prior probability, Pr(Ci) can be computed from Eq. (2) or Eq. (3):

$$\Pr(Ci) = \frac{\text{Total\_of\_Words\_in\_Ci}}{\text{Total\_of\_Words\_in\_Training\_Dataset}} \quad (2)$$

$$= \frac{\text{Size\_of\_Ci}}{\text{Size\_of\_Ci\_Training\_Dataset}} \quad (3)$$

Meanwhile, the evidence, which we call the normalizing constant of a particular word, wj, Pr(wj) is calculated by using Eq. (4):

$$\Pr(wj) = \frac{\sum \text{occurrence\_of\_wj\_in\_all\_categories}}{\sum \text{occurrence\_of\_all\_words\_in\_all\_categories}} \quad (4)$$

The total occurrence of a particular word in every category can be calculated by searching the training data base, which is composed from the list of word occurrences for every category [16]. As previously mentioned, the list of word occurrences for a category is generated from the analysis of all training documents in that particular category during the initial training stage. The same method can be used to retrieve the sum of occurrence of all words in every category in the training data base.

To calculate the likelihood of a particular category, Ci with respect to a particular word, wj, the lists of word occurrences from the training data base are searched to retrieve the occurrence of wj in Ci, and the sum of all words in Ci. These information will contribute to the value of Pr(wj|Ci) given in Eq. (5):

$$\Pr(wj|Ci) = \frac{occurrence\_of\_wj\_in\_Ci}{\sum \text{occurrence\_of\_all\_words\_in\_Ci}} \quad (5)$$

Based on the derived Bayes' formula for text classification, and the value of the prior probability Pr(Category), the likelihood Pr(Word|Category), and the evidence Pr(Word), along with the posterior probability, Pr(Category|Word) of each word in the input document being annotated to each category can be measured [4]. After all the posterior probabilities of each of the words in a particular document being annotated to each category have been computed, the overall probability for an input document to be annotated to a particular category, Ci is calculated by using the formula which is shown in Eq. (6):

$$\Pr(Ci|Document) = \frac{\Pr(Ci|w1,w2.....wn-1,wn)}{n} \quad (6)$$

where w1, w2. . …. wn _ 1, wn, are the words which are extracted from the input document.

The conventional Bayesian classifier is able to determine the right category of an input document by referring to the highest probability value calculated by the trained classifier based on Bayes formula [13]. The right category is represented by the category which has the highest posterior probability value, Pr(Category|Document), as stated in Bayes classification rule.

B. 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] combining the HMM and SVM where HMMs are used to as a feature extractor and then a new feature vector is normalized as the input of SVMs, so the trained SVMs can classify unknown texts successfully, also by combing with Bayes [19] use to reduce number of feature which as reducing number of dimension .SVM is more capable [8] to solve the multi-label class classification.

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 labelled 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 [32] presents method which can handle numeric and categorical data.

New method is proposing [20] as FDT to handle the multi-label document witch reduce cost of induction, and [28] 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 [21] 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.

D. Decision Rule

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories [29]. A popular format for interpretable solutions is the disjunctive normal form (DNF) model. [30] A classifier for category ci built by an inductive rule learning method consists of a disjunctive normal form (DNF) rule. [4]. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the size of rules set without affecting the performance of the classification The [31] presents a hybrid method of rule based processing and back-propagation neural networks for spam filtering.

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 discuss 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 need 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].

It put forward the novel improved TF-IDF approach for text classification, and will focus on this approach in the remainder of this paper, and will describe in detail the motivation, methodology, and implementation of the improved TF-IDF approach. The paper discusses and analyzes the relationship among confidence, support, recall and precision, and then presents the experimental results [36].

III. 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.

The proposed work can be explained in flow diagram shows in Figure 1. 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.

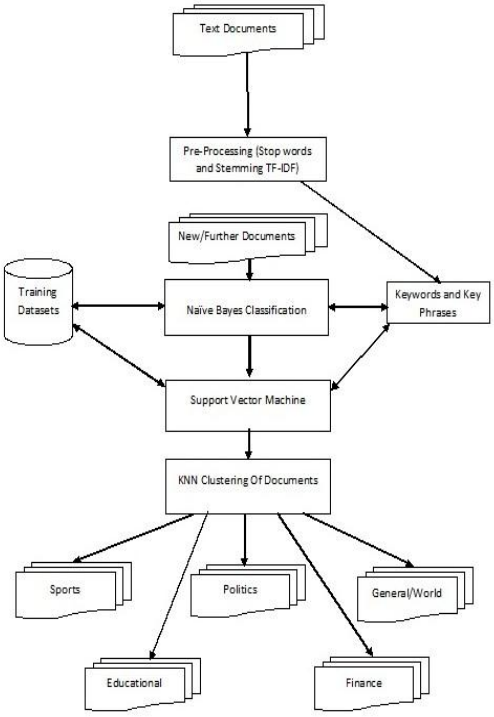

Figure 1: Proposed work

IV. APPLICATIONS

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.

V. 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 documents representation and an analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are proposed. This survey paper is also gives a brief introduction to the various text representation schemes. The existing classification methods are compared and contrasted based on various parameters namely criteria used for classification, classification time complexities and algorithms adopted. Different algorithms perform differently depending on data collection. To the certain extent SVM with term weighted VSM representation scheme performs well in many text classification tasks. In addition we add the keyword and key phrase extraction based classification to improve the time and accuracy of the document classification based on various features selection.

VI. 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] GongdeGuo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", Proc. ODBASE pp- 986 – 996, 2003
[10] EijiAramaki 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 QingyuanBai, "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 DaneshBehzadMoshiri "Improve text classification accuracy based on classifier fusion methods". 10th International Conference on Information Fusion, 1-6 2007.
[15] Sang-Burn Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng "Some Effective Techniques for Naïve Bayes Text Classification", IEEE Transactions on Knowledge and Data Engineering, Vol 18, NO 11, Nov 2006
[16] Nikos Tsimboukakis, and George Tambouratzis "Word-Map System for Content-Based Document Classification" IEEE Transaction on System, MAN, and Cybernetics, Vol. 41, NO. 5, 2011
[17] Lam Hong Lee, Dino Isa, WouOnnChoo, Wen YeenChue "High Relevance Keyword Extraction facility for Bayesian text classification domain of varying characteristic" Expert Systems with Applications 39 1147-1155, 2012
[18] L. Baker, A. McCallum. Distributional Clustering of Words for Text Classification, ACM SIGIR Conference, 1998
[19] R. Bekkerman, R. El-Yaniv, Y. Winter, N. Tishby. On Feature Distributional Clustering for Text Categorization. ACM SIGIR Conference, 2001.
[20] S. Basu, A. Banerjee, R. J. Mooney. Semi-supervised Clustering by Seeding. ICML Conference, 2002
[21] P. Bennett, S. Dumais, E. Horvitz. Probabilistic Combination of Text Classifiers using Reliability Indicators: Models and Results. ACM SIGIR Conference, 2002.

[22] P. Bennett, N. Nguyen. Refined experts: improving classification in large taxonomies. ACM SIGIR Conference, 2009
[23] S. Bhagat, G. Cormode, S. Muthukrishnan. Node Classification in Social Networks, Book Chapter in Social Network Data Analytics, Ed. Charu Aggarwal, Springer, 2011
[24] A. Blum, T. Mitchell. Combining labeled and unlabeled data with co-training. COLT, 1998
[25] D. Boley, M. Gini, R. Gross, E.-H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore. Partitioning-based clustering for web document categorization. Decision Support Systems, Vol. 27, pp. 329–341, 1999
[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] 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
[28] C.Apte, F. Damerau, and S.M. Weiss "Automated Learning of Decision Rules for TextCategorization", ACM Transactions on Information Systems, 1994
[29] Sholom M. Weiss NitinIndurkhya, "Rule-based Machine Learning Methods for Functional Prediction", Journal of Artificial Intelligence Research 3 383-403 1995
[30] 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
[31] 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
[32] 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
[33] Chen donghui Liu zhijing, "A new text categorization method based on HMM and SVM", IEEE 2010
[34] 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
[35] Dagan, I., Karov, Y., and Roth, D. "Mistake-Driven Learning in Text Categorization." In Proceedings of CoRR. 1997
[36] MIgual E .Ruiz, PadminiSrinivasn, "Automatic Text Categorization Using Neural networks", Advances in Classification Research, Volume VIII
[37] 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
[38] Hwee TOU Ng Wei Boon GohKok Leong Low, "Feature Selection, Perception Learning, and a Usability Case Study for Text Categorization", SIGIR 97 Philadelphia PA, 1999