Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Text Classification using Keyword Extraction Technique

Menaka S*

Research Scholar
PSGR Krishnammal College for Women
Peelamedu, Coimbatore
India

Radha N

Assistant Professor
GRG School of Applied Computer Technology
PSGR Krishnammal College for Women
Peelamedu, Coimbatore, India

Abstract— Text classification is the process of classifying the text documents based on words, phrases and word combinations with respect to set of predefined categories. Text classification has many applications such as mail routing, email filtering, content classification, news monitoring and narrow-casting. Keywords are extracted from documents to classify the documents. Keywords are subset of words that contains the most important information about the content of the document. Keyword extraction is a process used to take out the important keywords from documents. In this proposed system keywords are extracted from documents using TF-IDF and WordNet. TF-IDF algorithm is used to select the candidate words. WordNet is a lexical database of English which is used to find similarity among the candidate words. The words which have highest similarity are taken as keywords. The experiment has been done using Naive Bayes, Decision tree and K-Nearest Neighbor (KNN) algorithms and its performance are analyzed. Decision tree algorithm gives the better accuracy for text classification when compared to other algorithms.

Keywords— Classification, WordNet, TF-IDF, KNN, SOM, Lexical.

I. INTRODUCTION

Over the last decade, the number of digital documents available for various purposes has grown enormously with the increasing availability of high capacity storage hardware and powerful computing platforms. The vivid increase of documents demands effectual organizing and retrieval methods mainly for large documents. Text classification is one of the key techniques in text mining to categorize the documents in a supervised manner. The processing of text classification involves two main problems are the extraction of feature terms that become effective keywords in the training phase and then the actual classification of the document using these feature terms in the test phase. This text classification task has numerous applications such as automated indexing of scientific articles according to predefined thesauri of technical terms, routing of customer email in a customer service department, filing patents into patent directories, automated population of hierarchical catalogues of Web resources, selective dissemination of information to consumers, identification of document genre, or detection and identification of criminal activities for military, police, or secrete service environments and so on. Text classification can be used for document filtering and routing to topic-specific processing mechanisms such as information extraction and machine translation. Various methods are used for document classification such as Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Fuzzy C-means, Neural Networks, Decision trees and Rule based learning algorithms.

In the proposed work, the keywords are extracted from documents using TF-IDF and WordNet. There are limited number of words are selected from each document. Based on the extracted keywords, documents are classified using machine learning techniques. Section II describes the existing work done by different authors. Section III describes about the basic concepts of Wordnet. Section IV describes the keyword extraction and document classification process. Section V shows the experimental results of proposed work. Section VI shows the conclusion of the proposed work.

II. LITERATURE REVIEW

Wongkot Sriurai [1] compared the feature processing techniques of Bag of Words (BOW) with the topic model. Text categorization algorithms such as Naive Bayes (NB), Support Vector Machines (SVM) and Decision tree are used for experimentation. Then the results proved that the topic-model approach for representing the documents yield the best performance based on F1 measure of 79% an improvement of 11.1% over the BOW model. Hidayet Takci and Tunga Gungor [2] used centroid–based classification approach with inverse class frequency to classify the language independent text documents. The experiment is done on ECI/MC1 multilingual corpus and the results obtained shows better accuracy of 99% when compared to other methods. Xiaogang Peng and Ben Choi [3] proposed automatic classification of documents based on semantic hierarchy. The experiment is done on yahoo.com and yields results of 77.46%. Antonie and Zaialne [4] experimented text document categorization on Reuters-21578 using association rule mining. The experiment yields the micro-average of 81.8% values and the macro-average of 78.24%. The string kernel was proposed as the solution to the two main problems which is inherent in encoding documents into numerical vectors. Ercan and Cicekli [5] described a supervised learning approach that uses lexical chains for extraction. The main idea is to find

semantically similar terms, i.e., lexical chains, from text and utilize them for keyword extraction as semantic features. The experiment yield 45% precision for full text and 20% for abstract. Kardon et al. [6] discussed text mining algorithms to extract keywords from learning objects and also used WordNet to find semantic distance between the keywords. Then the keywords which have the highest similarity are selected as output keywords. Shoba et al. [7] proposed Kohonen's Self Organizing Map, an unsupervised learning technique for document classification. By using Kohonen's SOM, the dimensionality is reduced from a very high dimension data into 2 or 3 dimensional space. The kohonen's SOM algorithm is experimented with 7,000 files from 20 Newsgroup and gets the accuracy of 67% with mapping time as 91 seconds.

III. WORD NET

WordNet is a lexical database developed by George Miller at the Cognitive Science Laboratory at Princeton University. The basic building block of WordNet is synsets. Synsets is a collection of one or more synonyms. All synsets are classified into four parts are nouns, verbs, adjectives and adverbs. The organization of WordNet through lexical significances instead of using lexemes makes it different from the traditional dictionaries and thesaurus [8]. The difference between WordNet and other traditional dictionaries is the separation of the data into four databases associated with the categories of verbs, nouns, adjectives and adverbs.

In WordNet nouns are related to nouns only, verbs are related to verbs only, etc. The following list specifies the semantic relations available in WordNet.

Synonymy: Describes the relation binding between two equivalent or close concepts.
Antonymy: Represents the relation binding of two opposite concepts.
Hyperonymy: Semantic relation between more general words to more specific word
Meronymy: Denotes a part of something but which is used to refer to the whole of it.

IV. TEXT CLASSIFICATION

Text classification is one of the main applications of machine learning. The task is to assign unlabeled new text document to predefined category. The processing of text classification involves two main problems, first problem is the extraction of feature terms that become effective keywords in the training phase and then the second is actual classification of the document using these feature terms in the test phase. Before classifying documents, preprocessing has done. In preprocessing stop words are removed and the words are stemmed. Then the term frequency is calculated for each term in a document and also TF-IDF is calculated.

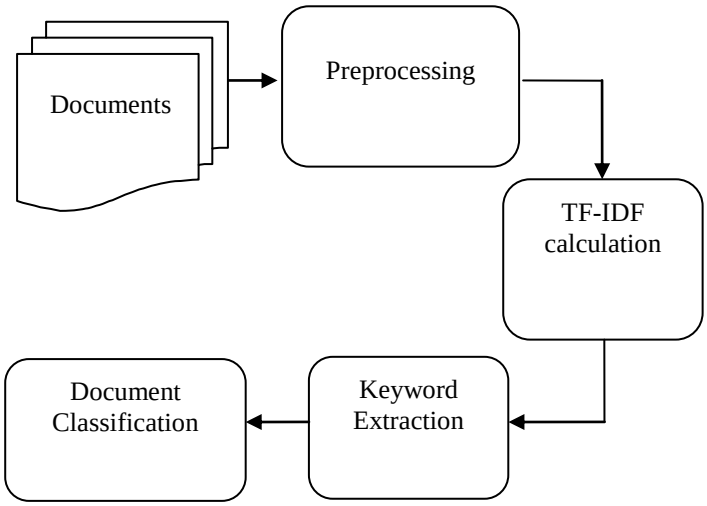

Fig. 1 Keyword based Document Classification

A. KEYWORD EXTRACTION

Keywords can be considered as condensed version of documents and short forms of their summaries. Keyword extraction is a significant technique for number of text mining related tasks such as document retrieval, webpage retrieval, document clustering and summarization. The main aim of keyword extraction is to extract the keywords with respect to their relevance in the text. First step is to select the desired documents (such as pdf files or text files) and it can be preprocessed.

Stop Words Elimination

Stop words are a part of natural language that do not have so much meaning in a retrieval system. The reason that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words are in text documents are prepositions, articles, and pro-nouns etc that does not provide the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc. Stop words are eliminated from documents because those words are not considered as keywords in text mining applications.

Stemming

Stemming techniques are used to find out the root/stem of a word. Stemming converts words to their stems which incorporates a great deal of language-dependent linguistic knowledge. For example, the words, connection, connects, connected, connecting all can be stemmed to the word 'connect'. In the present work, the Porter Stemmer algorithm is used which is the most commonly used algorithm in English.

Term Frequency-Inverse Document Frequency

Term Frequency–Inverse Document Frequency (tf-idf) is a numerical statistic which reveals that a word is how important to a document in a collection. Tf-idf is often used as a weighting factor in information retrieval and text mining. The value of tf-idf increases proportionally to the number of times a word appears in the document, but is counteracting by the frequency of the word in the corpus. This can helps to control the fact that some words are generally more common than others. Tf–idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

Tf–idf is the product of two statistics which are term frequency and inverse document frequency. To further distinguish them, the number of times each term occurs in each document is counted and sums them all together.

Term Frequency- Term Frequency (TF) is defined as number of times a term occurs in a document.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\text{maximum occurences of words}} \quad (1)$$

Inverse Document Frequency- An Inverse Document Frequency (IDF) is a statistical weight used for measuring the importance of a term in a text document collection. IDF feature is incorporated which reduces the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

$$idf(t,d) = \log \frac{|D|}{(\text{no. of documents term t appears})} \quad (2)$$

Then Term Frequency - Inverse document frequency [TF-IDF] is calculated for each word using the formula,

$$tfidf(t,d,D) = tf(t,d) \times idf(t,d) \quad (3)$$

In this equation (1) and (2) $f_{t,d}$ denotes the frequency of the occurrence of term *t* in document *d*. In equation (3) TF-IDF is calculated for each terms in the document by using Term Frequency ($tf_{t,d}$) and Inverse Document Frequency ($idf_{t,d}$).

Word Sense Disambiguation

Word sense disambiguation (WSD) is the activity of finding, given the occurrence in a text of an ambiguous word, i.e., polysemous or homonymous word, the sense of this particular word occurrence. For example, bank may have at least two different senses in English, as in the Bank of England denotes a financial institution or the bank of river Thames. WSD task is to decide that the occurrence of a word bank in the sentence 'Last week I borrowed some money from the bank' has which of the above senses. WSD is most important for numerous applications with natural language processing and indexing documents by word senses rather than by words for IR purposes. WSD may be seen as a text categorization task once word occurrence contexts can be viewed as documents and word senses as categories.

In this WordNet dictionary is used to find the concept similarity between words and word sense disambiguation. If the two words has the same sense then the word has the highest weight is selected as keyword. Based on this top keywords are selected and are stored in a database.

B. FEATURE SELECTION

Feature selection is a process commonly used in Machine Learning field to reduce the dimensionality of the feature space. The subset of the features available in the data is keywords are selected out. The selected features receive the highest scores according to a function that measures the importance of the feature for text classification task. The functions used to measure the importance are quite significant. Simple and effective function is the term frequency of a term that is only the terms that occur in the highest numbers in a document are retained and another one is tf-idf.

C. MACHINE LEARNING TECHNIQUES

The experiments are done using three different machine learning methods such as Naïve bayes, Support Vector Machine and k-Nearest Neighbour.

(i) *K-Nearest Neighbor:* The k-nearest neighbor algorithm (k-NN) is used to test the degree of similarity between documents and k training data. This method is an instant-based learning algorithm that categorized items based on closest feature space in the training set. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document. This method is non parametric, effective and easy for implementation.

(ii) *Naive Bayes Algorithm:* Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. The more expressive term for the underlying probability model would be independent feature model. This independence hypothesis of features make the features order is irrelevant and as a result that the presence of one feature does not affect other features in classification tasks which makes the computation of Bayesian classification approach more efficient. Naive Bayes classifiers can be trained powerfully by requiring a small amount of training data to estimate the parameters necessary for classification.

(iii) *Decision Tree:* The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In the decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories. The well-organized decision tree can easily classify a document by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf which represents the goal for the classification of the document. The decision tree classification method is outstanding from other decision support tools with several advantages. The main advantage of decision tree is its simplicity in understanding and interpreting, even for non-expert users. The major risk of implementing a decision tree is it over fits the training data with the occurrence of an alternative tree that categorizes the training data worse but would categorize the documents to be categorized better.

V. EXPERIMENTAL RESULTS

In this experiment is done by using journal papers and the classification is performed using an open source tool. RapidMiner is an environment for machine learning, predictive analytics, data mining, text mining, and business analytics. This tool is used for research, training, education, application development, rapid prototyping, and industrial applications. It provides data mining and machine learning procedures including data loading and transformation, data pre-processing and visualization, modeling, evaluation, and deployment. This tool is written in Java programming language, it uses learning schemes and attributes evaluators from the Weka machine learning environment and statistical modeling schemes from R-Project.

The proposed work is experimented by using 20 journal papers. Papers are collected manually from different journals. The efficient keywords from the journals are extracted using TF-IDF and WordNet. This keyword extraction process is developed in Java. Then the extracted keywords are stored in the database for classification. The documents are classified based on five predefined classes using machine learning algorithms. The five classes include finance, computer, mechanics, sports and medical. The 10 fold cross validation is used to evaluate the robustness of the classifiers. The prediction accuracy and the training time are two conditions used to evaluate the performances of the trained models and the prediction accuracy of the each model is compared. Table I shows the precision and recall values for Naive bayes classifier. The 10-fold cross validation results of the three classifiers Naive Bayes (NB), K-Nearest Neighbor (KNN) and Decision tree are summarized in Table IV

TABLE I
PRECISION AND RECALL FOR NAIVE BAYES CLASSIFIERS

Class	**Decision Tree**	
	Precision (in %)	Recall (in %)
C1	97.56	100
C2	97.50	97.50
C3	97.50	97.50
C4	100	100
C5	97.50	97.50

TABLE II
PRECISION AND RECALL FOR DECISION TREE CLASSIFIERS

Class	**Naive Bayes**	
	Precision (in %)	Recall (in %)
C1	90.48	95.00
C2	91.18	77.50
C3	82.22	92.50
C4	86.11	77.50
C5	86.05	92.50

TABLE III
PRECISION AND RECALL FOR KNN CLASSIFIERS

Class	**K-NN**	
	Precision (in %)	Recall (in %)
C1	100	100
C2	93.50	95.00
C3	95.12	97.50
C4	89.47	85.00
C5	88.37	95.00

TABLE IV
PERFORMANCE COMPARISON OF CLASSIFIERS

Criteria	**Naive Bayes**	**Decision Tree**	**KNN**
Accuracy	87.09	98.47	94.47
Absolute error	0.133	0.015	0.055
Root mean squared error	0.307	0.068	0.207
Root relative squared error	0.306	0.069	0.208

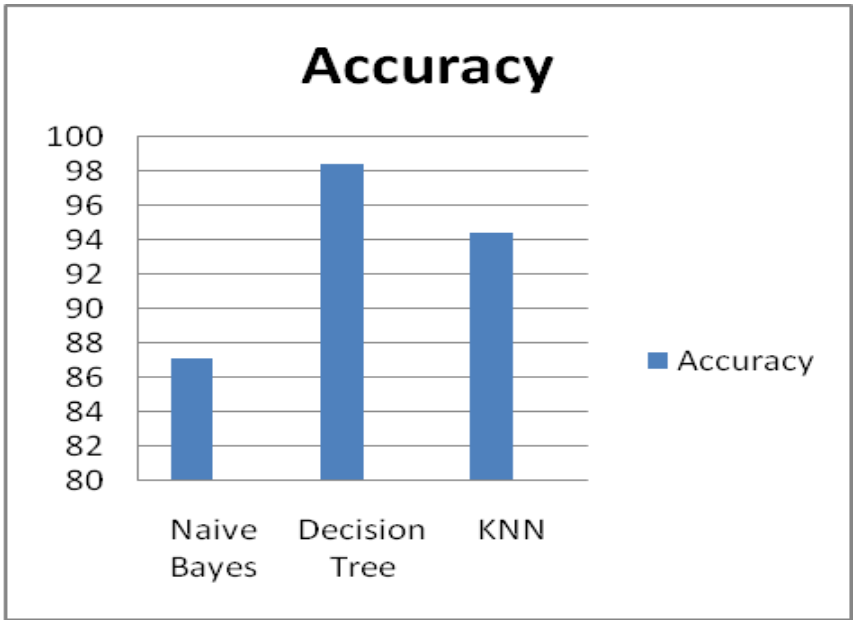

Fig.2 Prediction Accuracy

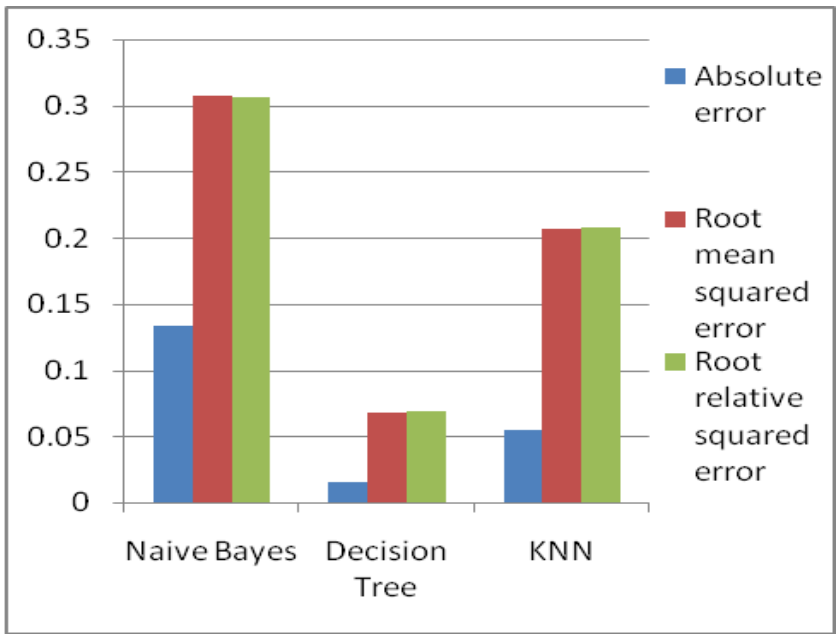

Fig. 3 Evaluation Criteria for Classifiers

VI. CONCLUSION

Text classification is one of the major applications of machine learning. The proposed method use text mining algorithms to extract keywords from journal papers. The WordNet dictionary is used to calculate the semantic distances between the keywords. The extracted keywords are having the highest similarity. Then documents are classified based on extracted keywords using the machine learning algorithms - Naïve Bayes, Decision Tree and k-Nearest Neighbor. The performance analysis of machine learning algorithms for text classification shows that the Decision Tree algorithm gives better results based on prediction accuracy when compared to other two algorithms.

REFERENCE

[1] Wongkot Sriurai, 2011, "Improving Text Categorization by Using a Topic Model" published in an International Journal Advanced Computing, Vol.2, Issue 6, pp 21-27.

[2] Hidayet Takci and Tunga Gungor, 2012, "A high performance centroid-based classification approach for language identification", published in Journal of Pattern Recognition Letters, Vol. 33, Issue 16, pp 2077-2084.

[3] Xiaogang Peng & Ben Choi, 2005, "Document Classifications Based on Word Semantic Hierarchies", in proceedings of ACM International conference on Artificial Intelligence and Applications, pp 362-367.

[4] Maria-Luiza Antonie, Osmar R. Zaiane, 2002 , "Text Document Categorization by Term association", in Proceedings of IEEE International Conference on Data Mining, pp 19-26.

[5] G. Ercan and I. Cicekli, 2007, "Using lexical chains for keyword extraction", published in International Journal of Information Processing and Management, Vol. 43, Issue 6, pp 1705–1714.

[6] Ahmad A .Kardan, Farzad Farahmandnia, Amin Omidvar, 2011, "A novel approach for keyword extraction in learning objects using text mining and WordNet", in proceedings of second World Conference on Information Technology, pp 788-792.

[7] B.H.ChandraShekar, Dr.G.Shoba, 2009, "Classifi-cation of Documents Using Kohonen's Self-Organizing Map", published in International Journal of Computer Theory and Engineering, Vol. 1, Issue 5, pp 1793-8201.

[8] A. Hulth. 2003 "Improved automatic keyword extraction given more linguistic knowledge", Proceedings of the 2003 conference on Empirical methods in natural language processing, p 216-223

[9] Jiyuan An, Yi-Ping Phoebe Chen, 2005 "Keyword Extraction for Text Categorization", in proceedings of International Conference on Active media Technology, p 556-561.

[10] O. W. Kwon and J. H. Lee, 2003, "*Text Categorization Based on K-nearest Neighbor Approach for Web Site Classification"*, published in an International Journal of Information Processing and Management, vol 39, p 25-44.

[11] Selamat and S. Omatu, 2004, "*Web Page Feature Selection and Classification Using Neural Networks"*, published in International journal of Information Sciences, vol 158, p 69-88.

[12] Show-Jane Yen, Yue-Shi Lee, Yu-Chieh Wu, jia-Ching Ying, Vincent S. Tseng, 2010, "*Automatic Chinese Text Classification using N-gram Model*", published in International journal on Computational Science and its Applications, vol 6018, p 458-471.

[13] Lingling Meng, Runqing Huang, Junzhong Gu, 2013, "*A Review of Semantic Similarity Measures in WordNet*", published in International Journal of Hybrid Information Technology, vol 6, issue 1, p 1-12.

[14] Yutaka Matsuo, Mitsuru Ishizuka. 2003 "*Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information*", published in International Journal on Artificial Intelligence Tools, vol 13, Issue 1, p 157-169.

[15] L. Baoli1, Y. Shiwen1, L. Qin2, 2003, "*An Improved k-Nearest Neighbor Algorithm for Text Categorization*", in Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, p 1503-1507.

[16] A. McCallum, and K. Nigam, 2003 "*A comparison of event models for naïve Bayes text classification*", published in Journal of Machine Learning Research, Vol. 3, p 1265–1287.

