A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms

Anuradha Patra
Barkatulallah University
Institute of Technology
Barkatulaah University Bhopal,
MP, India

Divakar Singh
Head CSE Deptt
Barkatullah University
Institute of Technology Barkatulaah
University Bhopal MP, India

ABSTRACT
Text classification approach gaining more importance because of the accessibility of large number of electronic documents from a variety of resource. Text categorization is the task of assigning predefined categories to documents. It is the method of finding interesting regularities in large textual, where interesting means non trivial, hidden, previously unknown and potentially useful. The goal of text mining is to enable users to extract information from textual resource and deals with operation such as retrieval, classification, clustering, data mining, natural language preprocessing and machine learning techniques together to classify different pattern. In text classification, term weighting methods design appropriate weights to the given terms to improve the text classification performance. This paper surveys of text classification, process of text classification different term weighing methods and comparisons between different classification algorithms.

General Terms
Data mining, text mining.

Keywords
Text categorization, natural language preprocessing, term weighing methods, classification algorithm.

1. INTRODUCTION
Text classification is an important part of text mining. Current research of text classification aims to improve the quality of text representation and develop high quality classifiers. Text classification process includes collection of data documents, data preprocessing, Indexing, term weighing methods, classification algorithms and fmeasure. Machine learning techniques have been actively explored for text classification. Among these are Naive bayes classifier [1], K-nearest neighbor classifiers [2], support vector machine [3], neural networks [4].

2. TEXT CLASSIFICATION PROCESS
Fig 1 represents the different stages of text classification which include collection of documents, preprocessing, feature indexing, feature filtering, different classification algorithm and performance measure

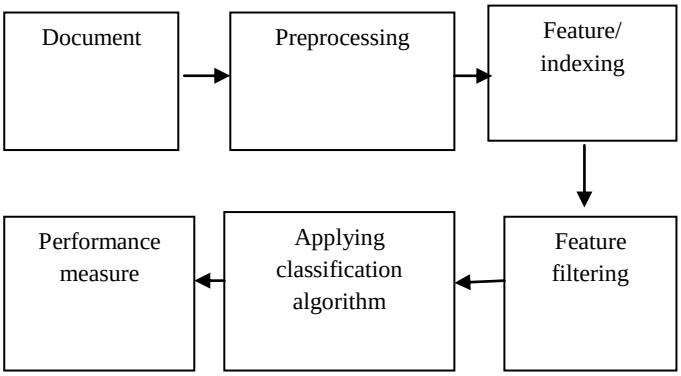

Fig 1. stages of text classification

2.1 Documents
First step is to collect the data from different type of format such as pdf, doc, html etc.

2.2 Preprocessing
Data mining is the process of extracting hidden pattern in a large dataset. real world data is often incomplete inconsistent and lacking in certain behavior and is likely to contain many errors. Data goes through a series of steps: Removing stop words: Stop words are very common words that appear in text that carry little meaning, they serve only syntactic meaning but do not indicate subject matter it is well recognized among the conformation retrieval experts that a set of functional English words (eg. “the”, “a”, “and”, “that”) is useless as indexing terms. These words have very low Discrimination value, since they occur in every English document [5]. Hence they do not help in distinguishing between documents with contents that are about different topics. The process of removing the set of non content- bearing functional words from the set of words produced by word extraction is known as stop words removal. In order to remove the stop words, this involves first creating a list of stop words to be removed, which is also called the stop word list. After this, the set of words produced by word extraction is then scanned so that every word appearing in the stop list is removed. Stemming: it is an approach used to reduce a word its stem or root from and is used widely in information retrieval tasks to increase the

recall rate[6] and give most relevant results such as: happy->happi,

3. INDEXING

The document has to be changed from the full text to a document vector. The most commonly used Document representation is called vector space model here documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix.VSM [7] representation scheme has its own disadvantages. Some of them are: high dimensionality of the representation, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document to overcome these problems, term weighting methods are used to assign appropriate weights to the term.

4. TERM WEIGHTING METHODS:

In text representation, terms are words, phrases, or any other indexing units used to recognize the contents of a text. each term in a document vector must be associated with a value called weight, which measures the importance of this term and denotes how much this term contributes to the categorization task of the document[8]. This section focuses on a number of traditional terms weighting methods. Table 1 comprise six different term weighing methods based on supervised and unsupervised methods

Table 1. Different term weighing methods

Methods	Term weighing factors	Denoted by	Description
Supervised term weighing methods	Chi square	χ^2	Multiply *tf* by χ^2 funtion
	Information gain	*ig*	Multiply *tf* by *ig* funtion
	Odd ratio	*OR*	Multiply *tf* by OR funtion
	Relevance factor	*rf*	Multiply *tf* by *rf* funtion
Unsupervised term weighing methods	Term frequency	*Tf*	Number of times term occur in adocuments
	Inverse document frequency	*Idf*	Multiply *tf* by idf funtion

Unsupervised methods don't have known information on category of training documents. TF(Term Frequency) and Idf(Inverse Document Frequency) are two main considerations of the traditional features weight algorithm. They have been widely used in information retrieval and better results have been achieved. tf refers to the number of times term appears in the text file. The features item can be a word, phrase, short language. Formerly, the document vectors were constructed based on TF. In different categories of documents, however, there may be big differences between the TF of feature items. Consequently, the frequency information is of importance to the text categorization [9]. IDF is the quantity of the feature item distribution in the document set. The Idf factor varies inversely with the number

$$n_i$$

of documents n_i which contain the term t_i in collection of N documents and is typically computed as $\log(\frac{N}{n_i})$. Supervised method is requiring known information on the category of training documents. Relevance factor can be calculated as $tf.rf = tf * \log(2 + \frac{a}{c})$. Where a is the number of document in the positive category that have that term. And c is the number of documents in the negative category that have this term Some researchers explain the term weighing methods by replacing idf factor with information gain, odds ratio, χ^2 [10],[11].

5. CLASSIFICATION ALGORITHMS

Databases are rich with hidden information that can be used for intelligent decision making. Classification algorithms can be used to extract models describing important data classes. Documents can be classified as supervised, unsupervised and semi supervised methods. There are several methods used to classify text such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Naive Bayes Classifier, and Decision Trees. Some techniques are described in sub section 5.

5.1 K-nearest neighbor classifiers

KNN is a classification algorithm where objects are classified by voting several labeled training examples with their smallest distance from each object. It was first describe in 1950 but initially applied to classification of news articles by Massand et al. in 1992 [12]. Yang compared 12 approaches to text categorization with each other, and judged that KNN is one of recommendable approaches, in 1999 [13]. Sebastiani in 2002 [14] evaluated as a simple and competitive algorithm with Support Vector Machine for implementing text categorization systems. The Major disadvantage of KNN is that it uses all features in computing distance and costs very much time for classifying objects

5.2 Naïve bayes classifier

Another famous and traditional approach to text categorization is NB. It learns training examples in priori probability given unseen examples. Basic concept is to calculate the probability it classifies documents based on learn advance before given unseen examples of categories and probabilities that attribute values belong to categories. The assumption that attributes are independent of each other underlies on this approach. Even though this theory violates the fact that attributes are dependent on each other, its performance is feasible. In text categorization [15] For vectorization performance of Naïve Bayes is very poor when features are co related to each other it is used popularly not only for text categorization, but also for any other classification problems, since its learning is fast and simple.

5.3 Support vector machines

It is a method for classification of linear and non linear data. This algorithm uses non linear mapping to transform training data into higher dimension and then it search for linear optimal separating hyper plane. SVM optimizes the weights of the inner products of training examples and its input vector, called Lagrange multipliers, instead of those of its input vector, itself, as its learning process .It provides a compact description of the learned model. A major research goal is in

SVM is to improve the speed in training and testing so that it become feasible option for large data set. In 1998, it was initially applied to text categorization by Joachims [16]. He explains the SVM in text categorization by comparing it with KNN and NB. Drucker et al. used SVM for implementing a spam mail filtering system and then compared it with NB in implementing the system in 1999 [17]. They conclude empirically that SVM was the improved approach to spam mail filtering than NB. In 2000, Cristianini and Shawe-Taylor presented a case of applying SVM to text categorization in their textbook [18].

5.4 Neural network

Neural network is a set connected input and output units in which each connection has a weight associated with it. Neural network learning is also referred to as connectionist learning due to the connections between units. It involves long training process they require number of parameters for classification of categories. A back propagation neural network [4]is a multilayer, feed-forward neural network consisting of input layer, a hidden layer and an output layer. The neurons present in hidden layer and output layer have biases, which are connection from units whose activation function is always 1.the bias term is also act as weights. The inputs are sent to the BPNN and the output obtained from the net could be 0 0r 1or bipolar (-1, +1) In 1995, Wiener was initially applied to text categorization by [19]. He used data Reuter 21578 for evaluating the approach to text categorization and conclude that performance of back propagation is better than KNN . In 2002, Ruiz and Srinivasan applied continually back propagation to text categorization [20].

5.5 Decision Tree

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. 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. : decision trees are built by greedy search algorithms, guided by some heuristic that measures "impurity". Irrelevant attributes may affect badly the construction of a decision tree. Small variations in the data can imply that very different looking trees are generated. Table 2 demonstrates the comparision between different classification algorithm according to advantages and disadvantages of algorithm.

6. PERFORMANCE MEASURE

To evaluate performance of text classifier first calculates precision and recall. let the document relevant to a query is denoted as retrieved. The set of documents that are both relevant and retrieved is denoted as relevant $\cap$ retrieved, precision is the percentage of retrieved documents that are in fact relevant to the query (i.e. "correct" responses). It is defined as

$$precision = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{Retrieved\}|}$$

Recall: this is the percentage of document that is relevant to the documents that are relevant to the query and were in fact, retrieved. It is formally defined as

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

And F measure [6] is

$$\text{F measure} = \frac{\text{Recall x } precision}{(\text{Recall} + precision)/2}$$

7. CONCLUSION:

This paper survey on text classification. This survey focused on the existing literature and explored the documents representation and an analysis of feature selection methods and classification algorithms Term weighting is one of the most vital parts for construct a text classifier. Different term-weighting approaches, including unsupervised and supervised term weighting, have been intensively investigated by previous studies. This paper also gives a brief introduction to the various text representation schemes. The existing classification methods are compared based on pros and cons. From the above discussion it is understood that no single representation scheme and classifier can be mentioned as a general model for any application Different algorithms perform differently depending on data collection.

Table 2. Comparison between classification algorithms

Classification algorithm	Formula	Pros	Cons
KNN	Distance measured by Euclidean distance $dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$	Simple and effective and easy to implement.	Hard to find out the value of K, time cost is more
NB	Posterior probability: $P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$	Easy for implementation and computation.	very poor when features are co related to each other
SVM	$\min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \| w \|^2 - \sum_{i}^{n} \alpha_i [y_i (w.x_i - b) - 1] \right\}$	compact description of the learned model, more capable to solve multi-label classification	Training speed is slow
NN	$I_j = \sum_i w_{ij} O_i + \theta_j$ J is hidden layer net input $O_j = \frac{1}{1 + e^{-I_j}}$ Output unit	Provide better result in complex domain	Long training process
Decision Tree	Partition of data, which is a set of training tuples and their associated class labels:then by making set of candidate attributes select the attribute by attribute selection methods.	Simple even non expert user can understand	Irrelevant attributes may affect badly the construction of a decision tree

8. REFRENCES

[1] Taeho Jo. "Neural Network for text categorization" International Journal of Information Studies 2010

[2] Gonde Guo, Hui Wang, David Bell,Yaxin Bi and Kieran Greer "KNN Model Based Approach in classification. pp .986-996 , 2003

[3] Sebastiani, F.." Machine Learning in Automated Text Categorization", ACM Computing Survey. pp. 1-47, 2002

[4] S. N.Sivanandam, S. N. Deepa "Principles of Soft Computing"

[5] Ramasundram, S.P.Victor "text categorization by BackPropagation" ,Proc.Int'l journal of computer application. pp. 0975-8887, 2010

[6] Deepika Sharma. "Stemming Algorithms: A Comparative Study and their Analysis" International Journal of Applied Information Systems. September 2012

[7] Vandana Korde,C Namrata Mahender "Text classification and classifier:A survey" International Journal of Artificial Intelligence & Applications. 2012

[8] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. "supervised and traditional term weighing methods for automatic textcategorization" ieee transactions on pattern analysis and machine intelligence. 2009

[9] Xiaojun Quan, Wenyin Liu, and Bite Qiu May" term weighing schemes for question categorization" ieee transactions on pattern analysis and machine intelligence. 2012

[10] Z.H. Deng, S.W. Tang, D.Q. Yang, M. Zhang, L.Y. Li, and K.Q.Xie, , "A Comparative Study on Feature Weight in Text Categorization,"Proc. Asia-Pacific Web Conf. pp. 588-597, 2004.

[11] F. Debole and F. Sebastiani."Supervised Term Weighting for Automated Text Categorization," Proc. ACM Symp. Applied Computing. pp. 784-788, 2003

[12] Massand.B, Linoff. G, Waltz. D "Classifying News Stories using Memory based Reasoning", the Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval. pp. 59-65, 1992.

[13] Yang, YAn evaluation of statistical approaches to text categorization, Information Retrieval. pp67-88. 1999.

[14] Sebastiani.F, "Machine Learning in Automated

Text Categorization", ACM Computing Survey. pp.1-47, 2002

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

[16] Joachims, T. "Text Categorization with Support Vector Machines: Learning with many relevantfeatures" europeon conference on machine learning pp 143-151, 1998

[17] Drucker, H., Wu, D., Vapnik, V. N. Support Vector Machines for Spam Categorization, IEEE Transaction on Neural Networks, pp.1048-1054 1999

[18] Cristianini.N, Shawe Taylor J. "Support Vector Machines and Other Kernel-based Learning Methods", CambridgeUniversity Press. 2000.

[19] Wiener E. D. "A Neural Network Approach to Topic Spotting in Text", The Thesis of Master of University of Colorado 1995.

[20] Ruiz M.E, Srinivasan P. "Hierarchical Text Categorization Using Neural Networks", Information Retrieval, pp 87-118. 2002.

[21] 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.