TEXT DOCUMENT CLASSIFICATION: A REVIEW

Priyanka[1], Amita Arora[2]
[1]Student, [2]Assistant Professor

Abstract
As most information is stored as text in web, text document classification is considered to have a high commercial value. Text classification is classifying the documents according to predefined categories. Complexity of natural languages and the very high dimensionality of the feature space of documents have made this classification problem difficult. In this paper we have given the introduction of text classification, process of text classification, overview of the classifiers and compared some existing classifier on basis of few criteria like time principle, merits and demerits.
Keywords: Text classification, Text Representation, Classifiers, Feature Selection

Introduction

Document classification is the task of grouping documents into categories based upon their content. The expansion of the internet has resulted in significant increase of unstructured data generated and consumed. The Web documents contain rich textual information, but the rapid growth of the internet has made it increasingly difficult for users to locate the relevant information quickly on the Web. Document retrieval, categorization, routing and filtering systems are often based on text classification. User feedback provides a set of training examples with positive and negative labels. Text classification have many challenges and difficulties. First, it is difficult to capture high-level semantics and abstract concepts of natural languages just from a few key words. For instance, there are many ways to represent similar concepts (e.g. agent, softbot, robot, or bot) and the same word can represent different meanings (e.g. bank can be either related to a finance problem or a river). Second, high dimensionality (thousands of features) and variable length, content and quality are the characteristics of a huge number of documents on the Web. Some of the techniques that are employed for document classification include Naïve Bayes classifier, Support Vector Machine, Decision Trees, Neural Network etc. Text document classification have applications in a wide variety of domains as described below:

News filtering and Organization: Most of the news services today are electronic in nature in which a large volume of news articles are created every single day by the organizations. In such cases, it is difficult to organize the news articles manually. Therefore, automated methods can be very useful for news categorization in a variety of web portals. This application is also referred to as text filtering. **Document Organization and Retrieval:** A variety of supervised methods may be used for document organization in many domains. These include large digital libraries of documents, web collections, scientific literature etc. Hierarchically organized document collections can be particularly useful for browsing and retrieval.

Opinion Mining: Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review.

Email Classification and Spam Filtering: It is often desirable to classify email in order to determine either the subject or to determine junk email in an automated way. This is also referred to as spam filtering or email filtering. Text classification is defining a set of logical rules that define that how to classify documents under the given set of categories. For example, automatically label each incoming news story with a topic like “sports”, “politics”, or “art” etc. A classification task starts with a training set D = (d_1….. d_n) of documents that are already labelled with a class C1, C2 (e.g. sport, politics). The task is then to determine a classification model which

is able to assign the correct class to a new document d of the domain. Text classification has two types as single label and multi-label .Single label document belongs to only one class and multi label document may belong to more than one classes. **Text Classification Process**

The various stages of text classification [2] are as follows:

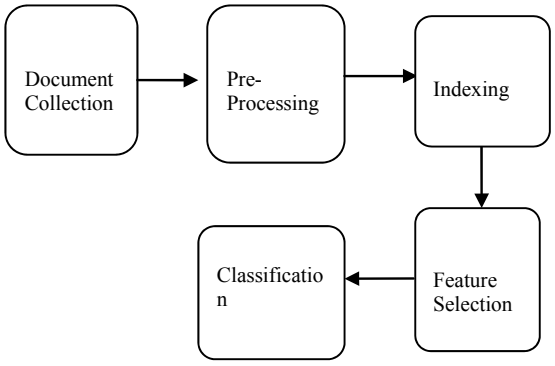

Figure 1. Document Classification Process

Documents Collection

This is first step of classification process in which we are collecting the different types (format) of document like html, .pdf, .doc, web content etc.

Pre-Processing

In pre-processing, following steps are involved: Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as "the", "a", "and" etc are frequent occurring, so these insignificant words need to be removed.

Stemming word*:* Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of reducing tokens to their root form, e.g. connection to connect, computing to compute etc.

Indexing

Indexing [3] plays an important role in text classification. Because documents cannot be directly interpreted by a classifier, they must be transformed into forms that a classifier can interpret through document indexing. The most commonly used document representation is called vector space model. In vector space model, documents are represented by vectors of words. Usually, one has a collection of documents which is represented by word by word document Matrix. BoW/VSM representation scheme has its own limitations. 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. But the major drawback of this model is that it results in a huge sparse matrix, which raises a problem of high dimensionality. Other various methods are presented in as 1) an ontology representation for a document to keep the semantic relationship between the terms in a document. 2) A sequence of symbols (byte, a character or a word) called N-Grams, which are extracted from a long string in a document. It is very difficult to decide the number of grams to be considered for effective document representation. 3) Multi-word terms as vector components. But this method requires a sophisticated automatic term extraction algorithms to extract the terms automatically from a document 4) Latent Semantic Indexing (LSI) which preserves the representative features for a document rather than discriminating features.

Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection. The main idea of Feature Selection [4] 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. In text classification a major problem is the high dimensionality of the feature space. There are various feature evaluation techniques such as information gain, term frequency, Chi-square, Gini index etc.

Document Frequency [5]: Document frequency is defined as the frequency of documents in which terms occur. For each term document frequency is calculated and if the document frequency of terms is less than some predetermined threshold then that term will be removed. Information Gain [6]: Information gain is the measure of how much information it contributes in the presence or absence of a term to make the classification decision on any class. Information gain reaches its maximum value when the document belongs to the respective class and the term is present in document.

Chi-square [6]: This is one of the most popular feature selection approaches. In statistics, the CHI2 test is used to examine independence of two events. In text feature selection, these two events correspond to occurrence of particular term and class, respectively. A high value indicates that the hypothesis of independence is not correct. If the two events are dependent, then the occurrence of the term makes the occurrence of the class more likely. Consequently, the regarding term is relevant as a feature. Chi-square score of a term is calculated for individual classes

Classification

The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Some machine learning approaches are Bayesian classifier, Decision Tree, K-nearest neighbor, Support Vector Machines, Neural Networks.

Classification Algorithms

The classifier classifies the feature vector to the category it belongs .Some of the classifiers are described below:

Rocchio's Algorithm

Rocchio's learning algorithm [7] is in the classical information retrieval. It was originally designed to use relevance feedback in querying full-text databases, Rocchio's Algorithm is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class c_i, and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

Ci= α * centroid ci − β * centroid ~ci. This algorithm is easy to implement, efficient in computation. The researchers have used a variation of Rocchio's algorithm in a machine learning context.

K-Nearest Neighbors

K-NN classifier [8] is a case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measures. This method is used for many applications 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. Modified traditional KNN use different K-values for different classes rather than fixed value for all classes

A major drawback of the similarity measure used in k-NN is that it uses all features in computing distances. Improved KNN algorithm [9] for text c1assification is based on particle swarm optimization which has the ability of random and directed global search within training document set. During the procedure for searching k nearest neighbors of the test sample, those document vectors that are impossible to be the k closest vectors are kicked out quickly. Besides it reduces the impact of individual particles from the overall. Moreover, the interference factor is introduced to avoid premature to find the k nearest neighbors of test samples quickly.

Naive Bayes

Naïve bias method is kind of module classifier under known priori probability and class conditional probability. Basic idea is to calculate the probability that document D is belongs to class C. There are two models are present for naive Bias as multivariate Bernoulli and multinomial model [10]. Out of these models multinomial model is more suitable when database is large. Poisson model for NB text classification and also give weight enhancing method to improve the performance of rare categories. Modified NB improves performance of text classification. Naïve Bayes is easy for implementation and computation. Performance of naïve bias is very poor when features are highly correlated and, it is sensitive to feature selection. In [11] naive bayes classifier has been used to classify the emotion of music based on lyrics.

Decision tree

When decision tree is used for text classification, it consists tree internal node are labeled as terms, branches departing from them are labeled by test on the weight, and leaf node are represented by

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, in [12] a method which can handle numeric and categorical data is used. New method is proposing [13] as "Fast decision-tree induction" to handle the multi-label document which reduces cost of induction. It uses a two strategy: (1) feature-set pre-selection, and (2) induction of several trees, each from a different data subset, with the combination of the results from multiple trees with a data-fusion technique tailored to domains with imbalanced classes.

Decision Rule

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories [14]. It constructs a rule set that describe the profile for each category. Rules are in the form of "If condition Then conclusion", where condition portion is filled by features of the category, and conclusion portion is represented with the categories name or another rule to be tested. This method [15] is capable to perform semantic analysis. The major drawback of this method is the need of involvement of human experts to construct or update the rule set.

Support Vector Machines

The support vector machines [15] need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed 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. The performance of the SVM classification remains unchanged even if documents that do not belong to the support vectors are removed from the set of training data. Amongst existing supervised learning algorithms, it is one of the most effective text classification methods [15] as it is able to manage large spaces of features and high generalization ability. But this makes this algorithm relatively more complex which in turn demands high time and memory consumptions during training stage and classification stage.

Neural Network

A neural network classifier [2] is a network of units, where the input units usually represent terms, the output unit represents the category. For classifying a test document, its term weights are assigned to the input units, the activation of these units is propagated forward through the network, and the value that the output units take up as a consequence determines the categorization decision. The single-layer perceptron is used due to its simplicity of implementation. The multi-layer perceptron which is more sophisticated, also widely implemented for classification tasks. In [16] the effectiveness of three neural networks, the Competitive, the Backpropagation and the Radial Basis Function, in text classification is examined and shown that Backpropogation and Radial Basis Function network outperform Competitive network because of the application of supervised learning.

Comparison Table

Name	Principle	Advantages	Disadvantages
Rocchio's Algorithm	$\mu = \frac{1}{\lvert D\rvert}\sum_{d\in D} V(d)$ where d is a document in D	Small computation	Low classification accuracy
KNN	Use similarity measures as criteria for classifying the documents $\cos(d_j, q) = \frac{\sum_{i=1}^{\lvert V\rvert} dq}{\sqrt{\sum_{i=1}^{\lvert V\rvert} d}}$	More local characteristic of document are considered as compared with rocchio's algorithm	Difficult in finding optimal value of K
Decision Tree	Entropy and information gain are used to for m decision tree	Reduce problem complexity	Once a mistake is made at higher level, subtree also become wrong

	Entropy = -p $\log_2 p$-q$\log_2 q$		
Support Vector Machine (SVM)	**The equation of a hyper plane is:** $W^T X + b = 0$	**Works well on numeric as well as textual data**	**Perform very poorly when features are highly co-related**
Neural Networks	**Pi=A.Xi Where X_i is the term frequency of ith document**	**Produce good result in complex domain**	**Training is relatively slow**

Conclusion

The text documents in digital form have drastically increased worldwide. Text classification has an important role in handling these documents. This review focuses on the existing literature and explored the documents representation and classification algorithms. Information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection. The existing classification methods are compared and contrasted based on various parameters. 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 with their respective advantages and disadvantages.

REFERENCES

[1] Y. H. Li and A. K. Jain, "Classification of Text Documents," http://citeseerx.ist.psu.edu /viewdoc/download?doi=10.1.1.100.7400

[2] Vandana Korde, C Namrata Mahender, "Text Classification and Classifiers: A Survey" , aircconline .com/ijaia/V3N2/3212ijaia08.pd

[3] Qi-RuiI Zhang , Ling Zhang , Shou-Bin Dong , Jing-Hua Tan, " Document Indexing In Text Categorization," http://ieeexplore.ieee.org/ abstract/document/1527600/

4] Foram P. Shah ,Vibha Patel, "A Review on Feature Selection and Feature Extraction for Text Classification" ieeexplore.ieee.org /iel7/7561562/7566075/07566545.pd

[5] Saket S. R. Mengle and Nazli Goharian, "Ambiguity Measure Feature -Selection Algorithm." Journal of the American Society for information Science and Technology, pp.1 037-1050

[6] Alper Kursat Uysal and Serkan Gunal, "A novel probabilistic feature selection method for text classification." https:// pdfs.semanticscholar.org/d151/00f63c40ae6a77 a6 e539b4c0d6b9f9f40b6f.pdf

[7]"Rocchioclassification",Nlp.stanford.http://nl p.stanford.edu/Ibook/html/htmledition/rocchio-classification-1 html.

[8] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", https://pdfs.semanticscholar.org/a7e2/814ec5db 800d2f8c4313fd436e9cf8273821.pdf

[9] Lingzhong Wang, Xia Li, "An improved KNN algorithm for text classification"http://ieeexplore.ieee.org/docume nt/5636476/

[10] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," https://link.springer.com/chapter/10.1007/978-1-4614-3223-4

[11] Yunjing An,Shutao Sun , Shujuan Wang , "Naive Bayes classifier for music emotion based on Lyrics" http://ieeexplore.ieee.org /abstract/document/7960070/

[12] Manish Mehta, Rakesh agrwal,"SLIQ: A Fast Scalable Classifier for Data Mining" http://sci2s.ugr.es/keel/pdf/algorithm/congreso/ SLIQ.pdf

[13] Peerapon Vateekul and Miroslav Kubat, "Fast Induction of Multiple Decision Trees in Text Categorization from Large Scale, Imbalanced, and Multi-label Data" http://ieeexplore.ieee.org/document/5360425/

[14] C.Apte, F. Damerau, and S.M. Weiss "Automated Learning of Decision Rules for Text Categorization", http://citeseerx .ist.psu.edu /viewdoc/ download ,doi=10.1.1.39.3129

[15] A.Khan, B.Baharudin, Lan Hong Lee, "A Review of Machine Learning Algorithms for Text Document Classification",http: //www .jait.us/uploadfile/2014/1223/201412230508005 32.pdf

[16] Zhan Wang, YifanHe ,Minghu Jiang , "Comparision among Three Neural Networks for Text Classification" http://ieee xplore.ieee.org/document/4129218/