International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 6 (July 2014) http://ijirae.com

Survey on Text Classification

K. Nalini
Research Scholar
Bharathiyar University
Coimbatore

Dr. L. Jaba Sheela
Professor, MCA Department
Panimalar Engineering College
Chennai

Abstract-The enormous amount of information stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific (pre-) processing methods and algorithms are required in order to extract useful patterns. Text Mining is the discovery of valuable, yet hidden, information from the text document. Text classification (Also called Text Categorization) is one of the important research issues in the field of text mining. It is necessary to classify/categorize large texts (documents) into specific classes. Text Classification assigns a text document to one of a set of predefined classes. This paper covers different text classification techniques and also includes Classifier Architecture and Text Classification Applications.

Key words: Text Classification, Preprocessing, Naïve Bayes Classifier, Association Based Classification, Decision Tree Induction.

I. Introduction

Text Mining [1] [2] refers to the process of deriving high-quality information from text. 'High quality' in text mining means that information extracted should be relevant to the user, and according to the interest of the user. Text mining [3] is similar to data mining, except that data mining tools [4] are designed to handle structured data from databases , but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. The corporate data is becoming double in size. Inorder the utilize that data for business needs, an automated approach is Text mining. By mining that text required knowledge can be retrieved which will be very useul. Knowledge from text usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization and predictive analytics. A typical application of text mining is to scan given set of documents written in a natural language and either to model them for predictive classification or populate a database or search index with the information extracted. Text (or Document) classification is an active research area of text mining, where the documents are classified into predefined classes.

Text Classification tasks can be broadly classified as Supervised Document Classification and Unsupervised Classification. In Supervised Document Classification some external mechanism (such as human feedback) provides information on the correct classification for documents or to define classes for the classifier, and in Unsupervised Document Classification (also known as document clustering), the classification must be done without any external reference and the system do not have predefined classes. There is also another task called Semi-Supervised Document Classification, where some documents are labeled by the external mechanism (means some documents are already classified for better learning of the classifier). 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 because to classifiy millions of text document manually is an expensive and time consuming task. In this paper we summarize text classification techniques that are used to classify the text documents into predefined classes [5].

A. Text Mining System – An Example

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant of knowledge for the user of that system. An example of Text Mining System Architecture is shown in figure 1.

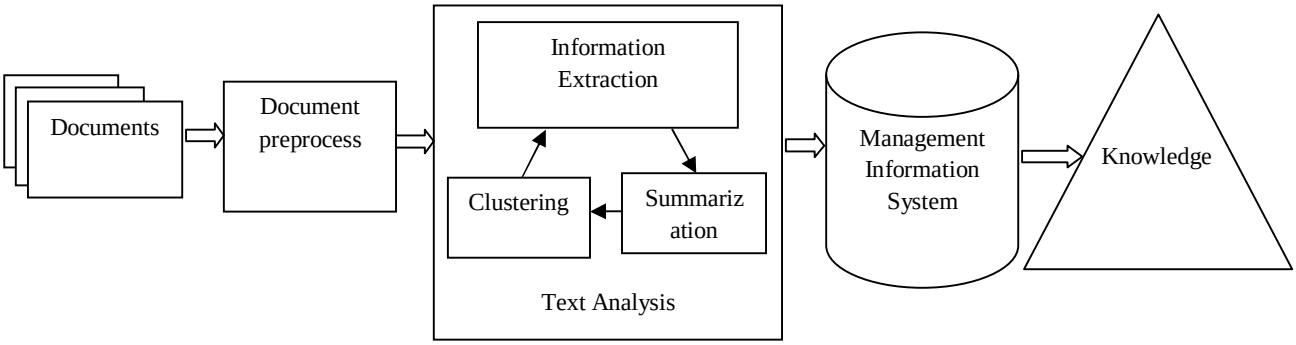

Figure.1 An example of Text Mining

B. Preprocessing

The main objective of pre-processing is to obtain the key features or key terms from stored text documents and to enhance the relevancy between word and document and the relevancy between word and category. Pre-Processing step is crucial in determining the quality of the next stage, that is, the classification stage. It is important to select the significant keywords that carry the meaning and discard the words that do not contribute to distinguishing between the documents. The pre-processing phase of the study converts the original textual data in a data mining ready structure. In general, text can be represented in two separate ways. The first is as a bag-of-words, in which a document is represented as a set of words, together with their associated frequency in the document. The bag-of-words model is a simplifying representation used in *natural language processing* and *information retrieval* . In this process, a text is represented as the container of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model is commonly used in methods of *document classification*, where the (frequency of) occurrence of each word is used as a *feature* for training a *classifier*. Such a representation is essentially independent of the sequence of words in the collection. The second method is to represent text directly as *strings*, in which each document is a sequence of words. Most text classification methods use the bag-of-words representation because of its simplicity for classification purposes.

The most common feature selection which is used in both supervised and unsupervised applications is that of stop-word removal and stemming. In stop-word removal, we determine the common words in the documents which are not specific or discriminatory to the different classes. A stop-words is a commonly occurring grammatical word that does tell us anything about documents content. Words such as ‘a’, ’an’, `the', `and', etc are stop-words. The process of stop-word removal is to examine documents content for stop-words and write any non-stop words to a temporary file for the document. We are then ready to perform stemming on that file. In stemming, different forms of the same word are consolidated into a single word. In this process we find out the root/stem of a word. The purpose of this method is to remove various suffixes, to reduce number of words, to have exactly matching stems, to save memory space and time. For example, singular, plural and different tenses are consolidated into a single word. We note that these methods are not specific to the case of the classification problem, and are often used in a variety of unsupervised applications such as clustering and indexing. In the case of the classification problem, it makes sense to supervise the feature selection process with the use of the class labels. This kind of selection process ensures that those features which are highly skewed towards the presence of a particular class label are picked for the learning process.

II. Classifier Architecture

Text classification is a fundamental task in document processing. The goal of text classification is to classify a set of documents into a fixed number of predefined categories/classes. A document may belong to more than one class. When classifying a document, a document is represented as a “bag of words”. It does not attempt to process the actual information as information extraction does. Rather, in simple text classification task, it only counts words (term frequency) that appear and, from the count, identifies the main topics that the document covers e.g. if in the document, cricket word comes frequently then “cricket” is assigned as its topic (or class) [6] [7]. There are two phases in the Classification. They are Training Phase and Testing phase.

A) Training phase

It is also called as Model Construction or Learning Phase; the set of documents used for model construction is called training set. It describes a set of predetermined classes. Each document/sample in the training set is assumed to belong to a predefined class (labeled documents). The model is represented as classification rules, decision trees, or mathematical formulae [7] [8].

B) Testing Phase

This is the 2nd step in classification. Also called Mode Usage or Classification Phase. It is used for classifying future or unlabeled documents. The known label of test document/sample is compared with the classified result to estimate the

accuracy of the classifier. For e.g. the labeled documents of the training set, is used further to classify unlabeled documents. Test set is independent of training set [6] [7] [8].

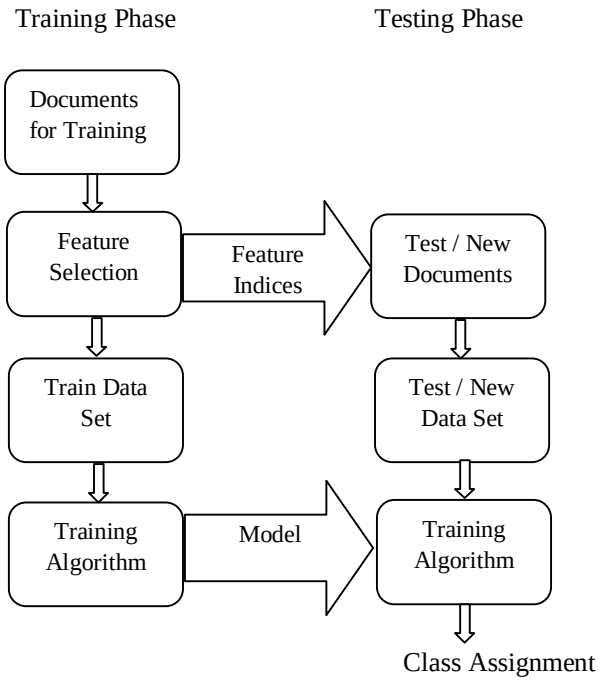

Figure 2: Flow Diagram of Text Classification

Using supervised learning algorithms [9], the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents). Figure 2 shows the overall flow diagram of the text classification task. Consider a set of labeled documents from a source D= [d1,d2,d3,…dn] belonging to a set of classes C = [c1,c2,c3,…cp]. The text classification task is to train the classifier using these labeled documents, and assign categories/classes to the new, unlabeled documents. In the training phase, the n documents are arranged in p separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process [6] [7]. Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection. In text manifold [19] categorization method, the text documents are treated as vectors in an n-dimensional space, where every dimension corresponds to a term. Then the metrics such as the cosine of the angle between two documents can be defined. However this space may be intrinsically located on the low dimensional manifold. The metric therefore should be defined according to the properties of manifold so as to improve the text categorization furthermore.

In TCBPLK [20] method of text categorization, texts are automatically assigned to appointed species according to text content. Similar texts are assigned to the same species through calculating the similarity among the texts. After the process of pattern aggregation for the word matrix, the numbers of words are greatly decreased, then TF.IDF method is applied to constructing the VSM. As the dimensions of the text are greatly decreased through the process of the P-L, the method decreases the learning time, and advances the speed and the of text categorization.

III. Text Mining Applications

The technology is now broadly applied for a wide variety of government, research and business needs. Some applications areas are Publishing and Media, Spam Filtering, Creating suggestion and recommendations, Fraud detection by investigation notification of claims, finding cyber bullying or cyber crime in IM and IRC chat.

A) Automated authorship attribution

Authorship attribution is the science of determining the author of a text document, from a predefined set of candidate authors or inferring the characteristic of the author from the characteristics of documents written by that author.

B) Automatic Document Distribution

Text classification also allows the efficient automatic distribution of documents via email or fax by eliminating the time consuming, manual process of faxing or mailing. And this can be achieved by first classifying the documents according to sender and message type.

C) Automated survey coding

Survey coding is the task of assigning a symbolic code from a predefined set of such codes to the answer that a person has given in response to an open-ended question in a questionnaire (survey). Survey coding has several applications, especially in the social sciences, ranging from the simple classification of respondents on the basis of their answers to the extraction of statistics on political opinions, health, and customer satisfaction etc.

D) Word sense disambiguation

Word sense disambiguation (WSD) is the activity of finding, given the occurrence in a text of an ambiguous (i.e. polysemous or homonymous) word, the sense of this particular word occurrence. For instance, bank may have (at least) two different senses in English, as in the Bank of England (a financial institution) or the bank of river Thames.

E) Text filtering

Text filtering is for example the activity to classify text document containing specific keywords or several keywords. Typical cases of filtering systems are e-mail filters, newsfeed filters.

F) Document clustering

Document clustering (also called unsupervised learning) is the act of collecting similar documents into clusters. In this, we do not have any external sources to provide information on the correct clustering for documents.

IV. Classification techniques

There are many techniques which are used for text classification. Following are some techniques:

- Association Based Classification
- Term Graph Model
- Centroid Based Classification
- Decision Tree Induction
- Classification using Neural Network

A) Association based Classification

Classification based on associations (CBA) [16] [17] [18] [19] [20] integrates classification and association rule mining. It generates class association rules and does classification more accurately than decision tree, C4.5. Classification association rules (CARs) are association rules with the class on the right hand side of the rules and conditions on the left side of the rules. These rules are extracted from the available training data and the most adequate rules are selected to build an "associative classification model".

1) Advantages: Associative classification has high classification accuracy and strong flexibility at handling textual data. The rules generated are also used for comparing the quality of different association rule mining approaches.

2) Limitations: CBA results in huge set of mined rules. It becomes challenging to store, retrieve, prune, and sort a large number of rules efficiently for classification.

B) Term Graph Model

Most existing text classification methods are based on representing the documents using the vector space model but due to this, sometimes, important information, such as the relationship among words, is lost. The term graph model [21] [22] [23] is an improved version of the vector space model. It represents not only the content of a document but also the relationship among words. A graph model is built to represent all extracted relations. To constructs graph, first, we construct a node for each unique term that appears at least once in the frequent datasets. Then we create edges between two nodes 'u' and 'v' if and only if they are both contained in one frequent dataset. The weight of the edge between 'u' and 'v' is the largest support value among all the frequent datasets that contains both of them.

1) Advantages: The graph representation of the document is more expressive than standard bag of words Representation, and consequently gives improved classification accuracy. We can also preserve and extract the hidden relationships among terms in the documents.

2) Limitations: The computational complexity of the graph representation for text classification is the main disadvantage of the approach.

C) Centroid Based Classification

Due to simplicity and linearity, for text classification Centroid Classifier [5] [24] [25] has become a commonly used method. Its basic idea is to construct a prototype vector, or centroid, per class using training documents. centroid-based classification algorithms are very fast, because only as many distance/similarity computations as the number of centroids

(i.e. classes) needs to be done. To improve the performance of centroid-based classifier normalization is an important factor when documents in text collection are of different sizes and/or the numbers of documents in classes are unbalanced. The classification task is to find the most similar class (with cosine similarity) to the vector of the document we would like to classify. Based on the documents in each class the centroid based classifier selects a single representative called centroid and then it works like KNN classifier with k=1.

1) Advantages: The technique is simple to implement and flexible to text data. It has relatively less computation than other methods in both the learning and classification stages.

2) Limitations: when a document from class A sharing more similarity with the centroid of class B than that of class A which lead to poor performance of the classification.

D) Decision Tree Induction

Decision trees [7] [8] are the most widely used inductive learning methods. Decision tree classification is the learning of decision trees from labeled training documents. One of the most well known decision tree algorithms is ID3 and its successor C4.5 and C5. A decision tree is a flowchart like tree structures, where each internal node denotes a test on document, each branch represents an outcome of the test, and each leaf node holds a class label. It is a top-down method which recursively constructs a decision tree classifier.

1) Advantages: Their robustness to noisy data and their capability to learn disjunctive expressions seem suitable for document classification. Decision trees are simple to understand and interpret.

2) Limitations: Decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where decisions are made at each node locally and cannot guarantee to return the globally optimal decision tree.

E) Classification using Neural Network

Neural networks [26] [27] [28] [29] have emerged as an important tool for classification. One major limitation of the statistical models (e.g. Naïve Bayes) is that they work well only when the underlying assumptions are satisfied. But neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Applications include bankruptcy prediction, handwriting recognition, speech recognition, fault detection, medical diagnosis etc. For classifying a given test document d_i, its term weights w_{ki} are loaded into the input units; the activation of these units is propagated forward through the network, and the value of the output unit(s) determines the categorization decision(s).

1) Advantages: Neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships. Neural networks are able to estimate the posterior probabilities, which provide the basis for establishing classification rule and performing statistical analysis. More than two hidden nodes provide better classification.

2) Limitation: With increase in the number of input and hidden nodes, the parameters needed for neural network also increases this result in over fitting of the data.

V. Conclusion

Text Classification is an important application area in text mining why because classifying millions of text document manually is an expensive and time consuming task. Therefore, automatic text classifier is constructed using pre classified sample documents whose accuracy and time efficiency is much better than manual text classification. If the input to the classifier is having less noisy data, we obtain efficient results. So during mining the text, efficient preprocessing algorithms must be chosen. The test data also should be preprocessed before classifying it. Text can be classified better by identifying patterns . Once patterns are identified we can classify given text or documents efficiently. Identifying efficient patterns also plays major role in text classification. Text classification techniques need to be designed to effectively manage large numbers of elements with varying frequencies. Almost all the known techniques for classification such as decision trees, rules, Bayes methods, nearest neighbor classifiers, SVM classifiers, and neural networks have been extended to the case of text data. Recently, a considerable amount of emphasis has been placed on linear classifiers such as neural networks and SVM classifiers, with the latter being particularly suited to the characteristics of text data. In recent years, the advancement of web and social network technologies have lead 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.

REFERENCES

[1] J.H. Kroeze, M.C. Matthee and T.J.D. Bothma**,** July 2007**,** "Differentiating between data-mining and text-mining Terminology..

[2] F. Sebastiani, 2002 "Machine learning in automated text categorization", ACM Computer Surveys 34(1), 1–47.

[3] Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.

[4] Navathe, Shamkant B., and Elmasri Ramez, (2000), "*Data arehousing And Data Mining*", in "*Fundamentals of Database System*

s", Pearson Education pvt Inc, Singapore, 841-872.
[5] Christoph Goller, Joachim Löning, Thilo Will and Werner Wolff, 2009, "Automatic Document Classification: A thorough Evaluation of various Methods", **"**doi=10.1.1.90.966".
[6] Vishal Gupta, Gurpreet S. Lehal, August 2009 "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1.
[7] Jiawei Han, Michelin Kamber, 2001, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, USA, 70-181.
[8] Megha Gupta, Naveen Aggrawal, 19-20 March 2010, "Classification Techniques Analysis", NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, pp. 128-131.
[9] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2005), "Experiments on Supervised Learning Algorithms for Text Categorization", International Conference , IEEE computer society, 1-8.
[10] B S Harish, D S Guru and S Manjunath, 2010, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR.
[11] Yu Wang and Zheng-Ou Wang, 2007, " A Fast KNN Algorithm for Text Classification", Machine Learning and Cybernetics, International Conference on, Vol. 6, pp. 3436-3441,Hong Kong, IEEE.
[12] Wei Wang, Sujian Li and Chen Wang, 2008, "ICL at NTCIR-7: An Improved KNN Algorithm for Text Categorization", Proceedings of NTCIR-7 Workshop Meeting, December 16–19, Tokyo, Japan.
[13] Jingnian Chen, Houkuan Huang, Shengfeng Tian and Youli Qu, 2009, "Feature selection for text classification with Naïve Bayes", Expert Systems with Applications: An International Journal, Volume 36 Issue 3, Elsevier.
[14] Wen Zhang, Taketoshi Yoshida and Xijin Tang, 2008, "Text classification based on multi-word with support vector machine", Journal: Knowledge Based Systems - KBS , vol. 21, no. 8, pp. 879-886, Elsevier.
[15] Steve R. Gunn, 1998, "Support Vector Machines for Classification and Regression", University of Southampton.
[16] Wenmin Li, Jiawei Han and Jian Pei, 2001, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules", IEEE International Conference on Data Mining - ICDM , pp. 369-376.
[17] Xiaoxin Yin, Jiawei Han. CPAR, 2003," Classification based on Predictive Association Rules", in Proceedings of SDM.
[18] Fernando Berzal, Juan-Carlos Cubero, Nicolás Marín, Daniel Sánchez, Jose-María Serrano, Amparo Vila, "Association rule evaluation for classification purposes".
[19] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, S. M. Kamruzzaman, May 2003, "Text Classification using the Concept of Association Rule of of data mining", International conference on Information on Information Technology, Kathmandu Nepal, pp234-241.
[20] Xin Lu, Barbara Di Eugenio, Stellan Ohlsson, 2007, "Learning Tutorial Rules Using Classification Based On Associations" , In Proceeding of the 2007 conference on Artificial Intelligence in Education,
[21] Wei Wang , Diep Bich Do , Xuemin Lin, 2005, "Term Graph Model for Text Classification",
[22] Chuntao Jiang, Frans Coenen, Robert Sanderson, Michele Zito, May 2010, "Text classification using graph mining-0based feature extraction", Journal Knowledge-Based Systems Volume 23 Issue 4, Elsevier.
[23] Dat Huynh, Dat Tran, Wanli Ma, Dharmendra Sharma, 2011, "A New Term Ranking Method Based on Relation Extraction and Graph Model for Text Classification", Faculty of Information Sciences and Engineering, University of Canberra ACT 2601, Australia.
[24] Songbo Tan, 2008, "An improved centroid classifier for text categorization", Expert Systems with Applications 35, 279–285, Elsevier.
[25] Verayuth Lertnattee, Thanaruk Theeramunkong, 2006, "Class normalization in centroid-based text categorization", Information Sciences 176, 1712–1738, Elsevier.
[26] Guoqiang Peter Zhang, November 2000, "Neural Networks for Classification: A Survey", IEEE Transactions on systems, man and cybernetics-Part C, Applications and Reviews, Vol. 30, NO. 4.
[27] Larry Manevitz, Malik Yousef, 2007, "One-class document classification via Neural Networks", Neurocomputing 70, 1466–1481, Elsevier.
[28] David Faraggi, Richard Simon, 1995, "The maximum likelihood neural network as a statistical classification model", Journal of Statistical Planning and Inference 46, 93-104, Elsevier.
[29] Ali Selamat, Sigeru Omatu, 2004, "Web page feature selection and classification using neural networks", Information Sciences 158, 69–88, Elsevier.
[30] Guihua Wen, Gan Chen, and Lijun Jiang (2006), "Performing Text Categorization on Manifold", 2006 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan , IEEE.
[31] JIAN-SUO XU (2007), "TCBPLK: A new method of text categorization ", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,, IEEE.

