IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

A Survey on Text Categorization

Senthil Kumar B[1], Bhavitha Varma E[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, India[1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, India[2]

Abstract: Now a day's managing a vast number of documents in digital forms is very important in text mining applications. Text categorization is a task of automatically sorting a set of documents into categories from a predefined set. A major characteristic or difficulty of text categorization is high dimensionality of feature space. The reduction of dimensionality by selecting new attributes which is subset of old attributes is known as feature selection. Feature-selection methods are discussed in this paper for reducing the dimensionality of the dataset by removing features that are considered irrelevant for the classification. In this paper we discuss several approaches of text categorization, feature selection methods and applications of text categorization.

Keywords: Text categorization, Clustering, Naïve Bayes, K Nearest Neighbor, Support Vector Machine.
.

I. INTRODUCTION

It is very difficult to apply techniques of data mining to textual data instead of numerical data. Therefore, it becomes necessary to develop techniques applied to textual data that are different from the numerical data. Instead of numerical data the mining of the textual data is called text mining. Text mining is procedure of synthesizing the information by analysing relations, the patterns and rules from the textual data. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. The functions of the text mining are text summarization, text categorization and text clustering. The content of this paper is restricted to text categorization.

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities. Automated prediction of trends and behaviours. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events. Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

The supervised learning algorithms still used to automatically classify text need sufficient documents to learn accurately while this proposed technique requires fewer documents for training. Here association rules from the significant words are used to derive feature set from pre-classified text documents. These rules are then used with the concept of Naïve Bayes classifier and during testing phase a concept of Genetic Algorithm has been added for final classification. Our observed experiment on this concept shows that the classifier builds this way is more accurate than the existing text classification systems.

II. FEATURE SELECTION

In machine learning and statistics feature selection also known as variable selection, attribute selection or variable subset selection. It is the process of detecting relevant features and removing irrelevant, redundant or noisy data.

A. Embedded Method

The embedded method incorporates feature selection process as a part of training process. They are specific to learning algorithms. It also captures feature dependencies. It uses independent criteria to decide optimal subsets for cardinality. Examples are classification trees, random forests and methods based on regularization technique. The inducer has its own FSA (either explicit or implicit). The methods to induce logical conjunctions provide an example of this embedding. Other traditional machine learning tools like decision trees or artificial neural networks are included in this scheme.

B. Wrapper Method

This method uses the predictive accuracy of a predetermined algorithm to determine the goodness of the selected subsets. Evaluation uses the criteria related to classification algorithm. Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive power. Wrapper methods based on SVM have been widely studied in machine-learning community. SVM-RFE (Support Vector Machine Recursive Feature Elimination), a wrapper method applied to cancer research is called, uses a backward feature elimination scheme to recursively remove insignificant features from subsets of features. In each recursive step, it ranks the features based on the amount of reduction in the objective function. It then eliminates the bottom ranked feature from the results. A number of variants also use the same backward feature elimination scheme and linear kernel.

C. Filter Method

This method is independent of learning algorithm. The filter method works well when the number of features is large. These methods select features based on discriminating criteria that are relatively independent of classification. Several methods use simple correlation coefficients similar to Fisher’s discriminant criterion. Others adopt mutual information or statistical tests (t-test, F-test). Earlier filter-based methods evaluated features in isolation and did not consider correlation between features. Recently, methods have been proposed to select features with minimum redundancy. The methods proposed use a minimum redundancy-maximum relevance (MRMR) feature selection framework. They supplement the maximum relevance criteria along with minimum redundancy criteria to choose additional features that are maximally dissimilar to already identified ones. By doing this, MRMR expands the representative power of the feature set and improves their generalization properties.

III.FEATURE REDUCTION

A. Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a gene selection procedure performed to decrease dimensionality of data. SVD is a matrix factorization method, and comes under linear vector algebra. In data analysis applied on gene expressions, the primary objectives of applying SVD are identification and extraction of the structural constitution within the data and also relating to significant associations involving gene expressions. The key method of SVD is to calculate the Eigen values and eigenvectors of the covariance matrix from the complete sample-gene matrix. The Eigen values or singular values will help deduce variation in the corresponding eigenvectors; If the singular value is greater, the respective eigenvector will contain a higher variability. Generally, some of the initially appearing eigenvectors that illustrate higher unpredictability are chosen as prime Principal Components (PCs) in order to reduce the data into a smaller quantity of dimensions. Since the remaining feature elements are removed, information is lost while the original data is recovered. This loss of data is utilized to obtain the feature genes.

B. Principal Components Analysis (PCA)

The PCA is a statistical data analysis method that transforms the initial set of variables into an assorted set of linear combinations, known as the principal components (PC), with specific properties with respect to variances. This condenses the dimensionality of the system while maintaining information on the variable connections [4]. The analysis is done on a data set by calculating and analysing the data covariance matrix, its Eigen values along with its respective eigenvectors systematized in descending order. Dimensionality Reduction is a process used in Data Mining where the numbers of random variables under consideration are reduced.

C. Independent Component Analysis (ICA)Independent component analysis (

D. ICA) is a computational technique used for splitting an assorted signal into its reduced subcomponents. A simple practice of ICA is the "cocktail party problem", wherein the fundamental speech signals are divided from a sample data comprising of individuals conversing together within a room. Usually this predicament is interpreted by considering the absence of time delays or echoes. An imperative note that is to be taken into consideration is that if N sources are present, at least N estimations (e.g. microphones) are required to mine the primal signals.

E. Canonical Correlation Analysis (CCA)

In statistics, Canonical Correlation Analysis, as presented by Harold Hostelling, is a method of creating sense out of cross-covariance matrices. If we take into consideration, two groups of variables and their correlations amongst the variables, then canonical correlation analysis will facilitate us to discover linear combinations of the variables and those variables which have the highest correlation with each other. CCA can also be used to create a model equation which relates two sets of variables, for instance a set of performance measures and a set of descriptive variables, or a set of outputs and a set of inputs.

F. Locally Linear Embedding (LLE)

Non-linear dimensionality reduction methods are largely categorized into two groups such asthe ones that offer a mapping from the high dimensional space to a lower dimensional embedding or vice versa, and the others that just provide a visualization in terms of graphs or charts in lower dimensionalities. In the perspective of machine

DOI 10.17148/IJARCCE.2016.5856

learning, mapping approaches may be regarded like an initial feature extraction step, which is followed by pattern recognition algorithms.

G. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a procedure used to overcome dimensionality reduction. It is used mainly in the Small Sample Size (SSS) problem. This issue can in a large set of data. In controlled experiments, the amount of existing cases is relatively rise in medical data sets where there are great amounts of dimensions or features less and the occurrence of features or variables is generally significantly larger than the size of the samples. At this point, the small sample size problem (SSS problem) arises.

IV. MACHINE LEARNING TECHNIQUES

Machine learning approach is used for classify the set of training data and automatically create the classifiers for the training data. Text classification is the task of automatically assign the texts into the predefined categories. In these text classification accomplished on the basis of endogenous collection of data. Text categorization mostly depends on the information retrieval technique such as indexing, inductive construction of classifiers and evaluation technique. In this machine learning, classifier learns how to classify the categories of documents based on the features extracted from the set of training data.Some of the key methods, which are commonly used for text classification, are as follows: neural network classifiers, support vector machines, Bayesian classifiers, boosting bagging classifiers.

A. Bayesian Classifier

The simple Bayesian classifier is mainly used for classification purpose. Using this, learn the profiles through the feedback collected from various websites. Based on the Bayesian classifier, the rank order of the pages will be rated. Learning the profile like LIBRA recommending system also uses the Bayesian classifier for the classification task. Text classification is based on calculating the posterior probability of the documents present in the different classes. Naive Bayes also a simple probabilistic classifier based on applying Bayes theorem with independence feature selection. Naïve Bayesian classification is used for anti-spam filtering technique. It has two different phases. The first phase has been applied for training set of data and the second phase employs the classification phase. Bayesian filter can also be used for classification of text. The result obtained by using this technique explains Bayesian approach is not sufficient for anti-spam filtering. Bayesian classifier can be applied on the reuters-21578; it improves the performance of the system.

B. Neural Network Classifier

Neural network classifiers are most important for the classification. The neural networks have the advantages of self-adaptive method. It means adjusting the weight themselves to the data without any specification and also it should be having the arbitrary accuracy. There are several types of neural networks are used for the classification task. There are feed forward multilayer networks and multilayerperceptron's are mostly used for neural network classifiers. Multilayer perceptron's have been applied successfully to solve many problems using the algorithm called Error back propagation algorithm. It has two passes. There are forward and backward pass. Feed forward back propagation neural network has the signal flow through the forward direction. Neural networks are very competitive to traditional classifiers for solving the classification problems. The basic unit of the neural network layer is neuron (or) unit. Each unit receives a set of inputs called Xi and associated set of weights W, corresponding to the term frequencies in the its document.

C. Support Vector Machines

The support vector machine method has been introduced in text classification by Joachim. Support vector Machines were used for separate the different classes. Linear support vector machines have the advantages of simplicity and interpretability. Normally, Text data which are correlated with one another and organized into the linearly separable categories. Support vector machines can be applied to the Email data classification. The performance of support vector machine is compared to the other classification techniques like decision trees, the rule based classifier and orchid method it should provide the more robust and flexible performance. The support vector machine classifier has been well suited for large amount of unlabeled data and small amount of labeled data. To solve the quadratic programming problem and two-class pattern recognition problem, support vector machine can be applied. Hyperplane are chosen for the separator for high dimensional surfaces. It should classify the positive and negative margins in the high dimensional surface. This method should not need any human and machines help for tuning on a validation set of parameters, default choices are available in the support vector machines. There is error estimating formulas are helpful for predicting the classification and eliminating the need of cross validation on the test and training set of data. It is very easy to select the features from the high dimensional space.

D. Decision Trees

Decision trees are designed essentially for hierarchical decomposition of the data space. Based on the attribute value it determines the predicate or condition. In this decision trees, class labels in the leaf node used for classification purpose. In order to reduce the over fitting data, pruning is to be done. There are several different kinds of splits in the decision trees are available. The listed splits are

- Single attribute split
- Similarity-based multi-attribute split
- Dimensional- based multi-attribute split

Decision tree implementations in the text context tend to be small variations compared to ID3, C4.5 for the purpose of adapting the text classification. The various algorithms which are used for the classification of data are decision

DOI 10.17148/IJARCCE.2016.5856

trees, linear programming, neural network and statistics. Among these algorithms Decision trees is one of the most popular and powerful approaches in data mining. The science and technology of exploring large and complex data to discover useful patterns this area is most importance for modelling and knowledge extraction from the data which are available. Both theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Decision trees, originally implemented in decision theory and statistics. The benefits of decision tree in data mining 1) It able to handle variety of input data such as nominal, numeric and textual. 2) It process the dataset that contain the errors and missing values. 3) It is available in in varies packages of data mining and number of platform.

E. k-Nearest Neighbor

The categorization itself is usually performed by comparing the category frequencies of the k nearest documents (neighbors).The evaluation of the closeness of documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. In the latter case the feature vectors have to be normalized to length 1to take into account that the size of the documents (and, thus, the length of the feature vectors) may differ. A doubtless advantage of the k-nearest neighbor method is its simplicity. It has reasonable similarity measures and does not need any resources for training. K nearest neighbor performs well even if the category-specific documents from more than one cluster because the category contains, e.g., more than one topic. This situation is badly suited for most categorization algorithms. A disadvantage is the above-average categorization time because no preliminary investment (in the sense of a learning phase) has-been done. Furthermore, with different numbers of training documents per category the risk increases that too many documents from a comparatively large category appear under the k nearest neighbors and thus lead to an inadequate categorization.

V. APPLICATIONS OF TEXT CATEGORIZATION

The applications of text categorization are manifold. Common traits among all of them are

- The need to handle and organize documents in which the textual component is either unique, or dominant, or simplest to interpret component.
- The need to handle and organize large quantities of such documents, i.e large enough that their manual organization into classes is either too expensive or not feasible within the time constraints imposed by the application.
- The fact that the set of categories is known in advance, and is variation over time is small.

VI.CONCLUSION

Text categorization play very important role ininformation retrieval, machine learning, text mining and it have been successful in tackling wide variety of real world applications. Key to this success have been the ever-increasing involvement of the machine learning community in text categorization, which has lately resulted in the use of the very latest machine learning technology within text categorization applications. Many approaches for text categorization are discussed in this paper. Feature selection methods are able to successfully reduce the problem of dimensionality in text categorization applications. Process of text classification is well researched, but still many improvements can be made both to the feature preparation and to the classification engine itself to optimize the classification performance for a specific application. Research describing what adjustments should be made in specific situations is common, but a more generic framework is lacking. Effects of specific adjustments are also not well researched outside the original area of application. Dueto these reasons, design of text classification systems is still more of an art than exact science.

REFERENCES

[1] Jaydeep Jalindar Patil, Nagaraju Bogiri, “Automatic Text Categorization Marathi Documents”, 2321-7782, International Journal of Advance Research in Computer Science and Management Studies, March-2015.
[2] Ashis Kumar Mandal, Rikta Sen, “Supervised Learning Methods for Bangla Web Document Categorization”, International Journal of Artificial Intelligence & Application (IJAIA),DOI:10.5121/ijaia.2014.5508 September 2014.
[3] Neha Dixit, Narayan Choudhary, “Automatic Classification of Hindi Verbs in Syntactic Perspective”, 2250-2459, International Journal of Emerging Technology and Advanced Engineering, August 2014.
[4] Aruna Devi, K., Saveetha, R., “A Novel Approach on Tamil Text Classification Using C-Feature”, 2321-0613, 2014. IJSRD International Journal of Scientific Research & Development, 2014.
[5] Nidhi, Vishal Gupta, “Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach”, DOI:10.5121/csit.2012.2421.
[6] Nidhi, Vishal Gupta, “Algorithm for Punjabi Text Classification”,0975-8887, International Journal of Computer Applications,January-2012.
[7] Nadimapalli V Ganapathi Raju et. al., “Automatic Information Collection & Text Classification for Telugu Corpus using K-NN”,2231-1009, International Journal of Research in Computer Application & Management, November-2011.
[8] K. Rajan et. al., “Automatic classification of Tamil documents using vector space model and artificial neural networks”, Expert Systems with Applications 36 (2009) 1091-10918, ELSEVIER, 2009.
[9] Abbas Raza Ali, Maliha Ijaz, “Urdu Text Classification”, FIT’09,December 16-18, 2009, CIIT, Abbottabad, Pakistan.[10] Munirul Mansur, Naushad UzZa man , Mumit Khan, “Analysis of N Gram Based Text Categorization for Bangla in Newspaper Corpus”.
[10] Kavi Narayan Murthy, “Automatic Categorization of Telugu News Articles”.
[11] Meera Patil, Pravin Game, “Comparison of Marathi Text Classifiers”, ACEEE Int. J. on Information Technology, DOI:01.IJIT.44.1., March 2014.
[12] István Pilászy, “Text Categorization and Support Vector Machines”.
[13] Bijal Dalwadi, Vishal Polara, Chintan Mahant, “A Review: Text Categorization for Indian Language”, 2349-4476, International Journal of Engineering Technology, Management and Applied Sciences, March 2015.
[14] Bhumika, Prof. Sukhjit Singh Sehra, Prof. Anand Nayyar, “AReview Paper on Algorithms Used for Text Categorization”, 2319-4847, International Journal of Application or Innovation in Engineering Technology & Management, March 2013.

DOI 10.17148/IJARCCE.2016.5856