Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 1; January-March, 2016, pp. 37-41

http://www.krishisanskriti.org/Publication.html

Comparative Survey of Document Analysis and Categorization Techniques

Sunita Rawat[1] and Manoj Chandak[2]

[1]Department of Information Technology GHRCE Nagpur, India
[2]Department of Computer Science and Engineering Ramdeobaba College of Engineering Nagpur, India

Abstract—*With an emergent quantity of existing online documents and the speedy expansion of the cyberspace, the job of automatic document categorization turned into the important technique for classifying the knowledge discovery and information. Appropriate classification of e-mails, e-documents, blogs, digital libraries and online news require machine learning, text mining and natural language processing techniques to obtain significant knowledge. The objective of this paper is to emphasize the significant methodologies and techniques that are in use in text document's classification. This paper presents an analysis of the techniques of document classification.*

Index Terms*: Documents classification, feature extraction, feature selection, text mining.*

1. INTRODUCTION

Due to the accessibility of the growing number of the documents on the internet from a diversity of sources, studies related to text mining are obtaining further significance. Researchers from several diverse fields attempt to utilize their own techniques to automatically organize these data collections and allow users to access data in some knowledgeable way. Classification is one of the techniques usually in use, which allows automatic steering of a meticulous document into some pre-specified class [1].

Now a days the internet is the key resource for the text documents; the quantity of textual data accessible to us is continuously increasing, as well as about 80 percent of the data of an association is stocked in amorphous textual format [3], like email, reports, news and views etc. In order to help out the analysis done by human, there is a requirement of automatically recovery of valuable information from the enormous quantity of textual data [6].

Market tendency depend on the details of the online articles, news, events and sentiments is a rising matter for research in text mining and data mining community. State-of-the-art methods to text classifications are given in [7], where they discussed about the classifier construction, documents representation and classifier evaluation. Document classification may be analyzed as transfer documents or document's portion in a predeclared group of classes. Generally this group is formed once hence it known as training documents, therefore, stays unaffected eventually.

2. DOCUMENT PREPROCESSING

Majority of the techniques applied in document classification also applied in data mining. In data mining the data which get examined is numerical, hence, before now in the representation needed by the algorithms.

Words of the documents have to translate into numerical forms to use these algorithms for document categorization. This process is known as document preprocessing, which consists of two steps, first is feature extraction, and second one is feature selection.

2.1 Feature Extraction

The common difficulty in this part is to create a list of terms that explains the documents adequately. Thus the parsing is done on training documents to find out a list of all words that is features available in the documents. Subsequently techniques known as feature reducing are used to decrease the dimension of the list formed by the parsing method; this list is known as dictionary. Stop word removal method and word stemming method are the main methods for this function.

The objective of stop word removal is to get rid of the dictionary from "noise" (e.g. prepositions, articles, numbers). By matching the entries of the dictionary with a predeclared stop word list, noise can be taken out from the document. Terms that vary merely in the affix are get treat by word stemming. Generally used word stemming methods are successor variety, affix removal, and n-grams [11].

2.2 Feature Selection

Subsequent to feature extraction, the next step is feature selection. Goal of this stage is to remove those features that give information which is not that much important. To find out the significant features, statistical values are utilized. The essentially used terms are TF, IDF and their product (TFxIDF), where TF stands for term frequency and IDF stands for inverse document frequency.

TF indicates that as compare to non significant words, significant words arise more frequently in a document. Whereas IDF indicates the uncommon words in the document collection are hypothetical to have the largest descriptive influence. By using TFxIDF the two terms are combined into one term. Finally top n words with the maximum score are chosen as features [10].

3. MACHINE LEARNING TECHNIQUES

By using supervised, semi supervised and unsupervised methods, the documents can be classified. For the clustering and categorization of online documents many algorithms and approaches are introduced newly.

In this paper we targeted on the supervised categorization approaches. Usually for automatic text categorization, supervised learning approaches are utilized, where pre-defined group labels are allocated to documents depending on the possibility recommended by a training set of labeled documents. Few of these approaches are discussed below.

3.1 Decision Trees

Decision tree approaches restructure the manual classification of the training documents via building precise true/false-queries similar to a tree structure where the nodes indicate questions and the leaves equivalent class of documents [12]. Subsequent to build the tree, a novel document can simply be classified by placing it in the tree's root node and allow it to run during the query structure awaiting it achieves an assured leaf.

The most important advantage of decision trees is the reality that the output tree is simple to infer even for people who are unknown with the particulars of the model.

The disadvantage of the decision tree method is "overfitting" [12].

3.2 Decision Rules

Decision rules categorization technique makes use of the rule based inference to categorize documents to their interpreted groups. A rule set is constructed by the algorithm that describes the outline for every one group.

Usually, a rule comprises of a group name with a dictionary feature. Next the rule set is formed with the logical operator "or" for merging the different rules. Generally all rules are not needed to classify the documents efficiently. Thus, to decrease the rule sets size, heuristics are used [13].

The major advantage of decision rules is the likelihood to produce local dictionaries throughout the feature extraction stage.

A disadvantage is that it is unfeasible to allocate a document entirely to one class since rules from various classes are related.

3.3 K-nearest Neighbor

This method totally omits the learning stage and classifies on-the-fly. The classification itself is typically done by matching the class frequencies of the k nearest neighbor (documents). Closeness among documents can be calculated by finding the angle among the two feature vectors or manipulating the Euclidean distance among the vectors.

Advantage of this method is it has realistic similarity measures and training resources are not required. This method does well even in managing the categorization jobs with multi-categorized documents.

The main disadvantage of this technique is while calculating distance it utilizes all features and motivates the technique computationally rigorous, mainly when the content of training set increases.

3.4 Bayesian Approaches

Naïve and non-naïve Bayesian methods are the two groups in document classification. A further expressive term for the fundamental probability model probably independent feature model. Due to this supposition the computation of Bayesian approaches becomes more competent. However, the supposition is obviously strictly violated in all language, according to studies; the categorization accurateness is not critically pretentious by this type of infringements [14]. Yet, numerous non-naïve Bayesian methods remove this supposition.

Advantage of these classifiers is that it needs a small quantity of training data to guess the constraints essential for classification and as well illustrate a great runtime-behavior throughout the categorization of novel documents [15]. However, a disadvantage of Bayesian methods normally is that they can merely process binary feature vectors [16], therefore, have to dump probably appropriate information.

3.5 Neural Networks

For document classification complications various neural network methods have been used. Whereas a few of them utilize perceptrons which is the easiest kind of neural networks, which made up of only two layers, one is input layer and other one is output layer [17], further difficult neural networks are made with a hidden layer between the input and output layers [18]. Usually, these feed-forward-networks made of as a minimum three layers in addition to use backpropagation as learning method.

Neural networks advantage is that they can manage noisy information superbly [19]. The benefit of the high suppleness of neural networks involves the disadvantage of excessive computing expenses. A further drawback is that neural networks are very hard to know for a normal customer.

3.6 Regression-based Methods

In this approach the training data are denoted as a couple of input-output matrices wherever the input matrix is similar to feature matrix A in addition to the output matrix B forms of

flags representing the group association of the equivalent document in matrix A. Hence matrix B has the identical amount of rows similar to A along with c columns where c denotes the sum of classes represented. The objective of the technique is to discover a matrix F that converts A into B' (by calculating B'=A*F) in order that B' equals B. The matrix F is calculated by using multivariate regression approaches [20].

An advantage of this technique is that morphological preprocessing of the documents can be passed up without loosing classification excellence. Regrettably regression-based approaches are not so trendy in the classification society; hence, study matching regression-based approaches with others are comparatively unusual.

3.7 Vector –based Methods

Support vector machines and centroid algorithm are the two vector-based methods which we discussed here.

Centroid algorithm is one of the easiest classification processes. Throughout the learning phase normal feature vector for every one class is evaluated and place as centroid vector for the class [21]. A novel document is effortlessly classified by getting the centroid vector nearest to its feature vector. This method does not need many training documents, unless the document clusters overlie everyone. Even if, the class forms of two or more dissimilar topics, the algorithm achieves regularly deprived. The technique is furthermore unsuitable if the amount of classes is huge [22].

Support vector machines (SVMs) are one of the discerning categorization techniques which are usually known to be more perfect. It is depend on the Structural Risk Minimization theory. The plan of this theory is to get a proposition to promise the minimum true error. Both positive and negative training set are needed by the SVM that are not common for further classification methods. These positive and negative training set are required to look for the decision surface for the SVM that most excellently divides the positive from the negative data in the n dimensional space. The concert of the SVM categorization stays unaffected if documents that do not fit into the support vectors are separated from the training data set.

An advantage of support vector machines is its better runtime-behavior throughout the classification of novel documents since only one dot product per novel document has to be evaluated.

A drawback is the reality that a document could be assigned to numerous classes since the resemblance is usually evaluated independently for every one class. However, for document categorization SVM is an excellent approach.

4. COMPARISON OF CATEGORIZATION METHODS

As shown, many algorithms have been proposed for document categorization. Table I gives an overview on all discussed algorithms. The studies utilize manually pre-categorized documents as input data set. Usually this data set is split in training and a test part; the latter is needed to determine the quality of the algorithm developed.

Commonly used test collections are the Reuters collection (newswires from Reuters; downloadable for instance at http://www.research.att.com/~lewis/reuters21578.html) and the OHSUMED collection (abstracts from medical journals). The comparison of algorithms and heuristics is scientifically demanding [23]. The results may be heavily dependent of the test data set.

Table 1: Comparison of Categorization Methods

Authors	Type of methods investigated								Test corpus			Main results
	Dec. Tree	Dec. Rules	k-NN	Bayes. Appr.	Neural Netw.	Regr.-based	Centroid	SVM	REUTERS	OHSUMED	Others	
[13]	x	x		x					x		x	Decision Rule shows the best performance, Decision Tree and Bayesian Independence Classifier perform similar but worse.
[2]	x			x					x		x	Both algorithms perform similar.
[4]			X	x			x				x	Performance of the algorithms is similar however the combination of the algorithms produce better results.

[22]	x			x			x	x	x			Support Vector Machines delivers the most excellent performance, while Similar (Centroid) the worst one.
[21]	x		X	x			x	x	x			SVM has the most excellent performance.
[5]			X				x		x	x		Combinations of ExpNet (k-NN) and Rocchio (Centroid) or ExpNet and Widrow-Hoff (Centroid) perform better than the basic algorithms.
[18]					x		x			x		Neural Networks perform better than Rocchio (Centroid).
[8]			X	x	x	x		x	x	x		With few documents per category (< 10), Support Vector Machines, k-NN, and LLSF (regression-based) perform significantly better than the other approaches; while, with greater than 300 documents per class performance of all the methods is similar.
[9]			X					x			x	SVM performs superior than k-NN.

Furthermore, several parameters usually have to be defined to initialize the procedures and the performance may depend on their initialization.

Taken these limitations into consideration, we have to emphasize that the SVM method has outperformed the other methods in several comparisons. Furthermore, there are results indicating that combinations of basics methods often provide better results than the application of the underlying "pure" methods.

5. CONCLUSION

This paper provides a review on feature extraction process, feature selection process and classification algorithms. Several algorithms or combination of algorithms as hybrid approaches was proposed for the automatic classification of documents, among these algorithms, SVM, NB and KNN classifiers are shown most appropriate in the existing literature.

REFERENCES

[1] Rennie and Jifile "An application of machine learning to e-mail filtering", KDD-2000 Workshop on Text Mining, Boston, 2000.

[2] Lewis and D. D. Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization", in Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pp 81-93.

[3] Raghavan, P., S. Amer-Yahia and L. Gravano eds., "Structurein Text: Extraction and Exploitation." In Proceeding of the 7th international Workshop on the Web and Databases (WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2004.

[4] Larkey L. S. and Croft W. B., "Combining Classifiers in Text Categorization", in Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 289-297.

[5] Lam W., and Ho C. Y., " Using a Generalized Instance Set for Automatic Text Categorization", in Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 81-89.

[6] Pegah Falinouss, "Stock Trend Prediction using News Article's:a text mining approach" Master thesis -2007.

[7] Sebastiani F., "Machine learning in automated text categorization" ACM Computing Surveys (CSUR) 34, pp.1 – 47, 2002.

[8] Yang, Y. and Liu X., "A Re-Examination of Text Categorization Methods", in Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 42-49.

[9] Siolas G. and d'Alché-Buc F., "Support Vector Machines based on a semantic kernel for text categorization", in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00), 2000, pp. 205-209.

[10] Jj Wei, C. P and Dong Y. X., "A Mining-Based Category Evolution Approach to Managing Online Document Categories", in Proceedings of the 34th Annual Hawaii International Conference on System Sciences, 2001.

[11] Bowman M., "Text Processing", URL: http://www.cse.ogi.edu/class/cse580ir/handouts/23%20September/Text%20 Processing [2001-09-26].

[12] Dd Gerstl, P., Hertweck, M. and Kuhn B., "Text Mining: Grundlagen, Verfahren und Anwendungen", in: Praxis der Wirtschaftsinformatik- Business Intelligence, 2001, Vol. 39, No. 222, pp. 38-48.

[13] Apté, C., Damerau, F. and Weiss S. M., "Towards Language Inde-pendent Automated Learning of Text Categorization Models", in Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Develop-ment in Information Retrieval, 1994, pp. 23-30.

[14] Domingos, P. and Pazzani M., "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", in: Machine Learning, 1997, Vol. 29, No. 2-3, pp. 103-130.

[15] Herrmann K., "Rakesh Agrawal: Athena: Mining-based Interactive Management of Text Databases", URL: http://www3.informatik.tu-muenchen.de/lehre/WS2001/HSEM-bayer/textmining.pdf [as of 2002-03-02].

[16] Cc Lam, W., Low, K. F. and Ho C. Y., "Using a Bayesian Network Induction Approach for Text Categorization", in: Proceedings of the 15th Inter-national Joint Conference on Artificial Intelligence, 1997, pp. 745-750.

[17] Ng, H. T., Goh W. B. and Low K. L., "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization", in Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Develop-ment in Information Retrieval, . 1997, pp. 67-73.

[18] Ruiz, M. E. and Srinivasan P., "Automatic Text Categorization Using Neural Network", in: Proceedings of the 8th ASIS SIG/CR Workshop on Classifica-tion Research, 1998, pp. 59-72.

[19] Krahl, D., Windheuser, U. and Zick, F.-K., "Data Mining – Einsatz", in der Praxis, Addison Wesley Longman: Bonn .

[20] Yang, Y. and Chute C., " An Example-Based Mapping Method for Text Categorization and Retrieval", in: ACM Transactions on Information Systems, 1994, Vol. 12, No. 3, pp. 253-277.

[21] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", in: Proceedings of the 10th European Conference on Machine Learning, 1998, pp. 137-142.

[22] Dumais, S. Platt, J., Heckermann, D. and Sahami M., "Inductive Learning Algorithms and Representations for Text", in: Proceedings of the 7th In-ternational Conference on Information and Knowledge Management, 1998, pp. 148-155.

[23] Golden, B. L. andStewart W. R., "Empirical Analysis of Heuristics", in: Lawler, E. L., Lenstra, J. K., Rinooy Kan, A. H. G., Shmoys, D. B. (Eds), The Traveling Salesman Problem, Wiley: New York, 1985, pp. 207-249.