A Review of Document Classification Techniques

[1] Harsimran Pal Kaur, [2] Kamaldeep Kaur
[1]PG Student,[2]Assistant Professor
Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India,

Abstract: — Now-a-days, vast amount of data is generated daily on the web. Data on the web is in the form of e-journals, e-newspapers, web pages etc. The most of the data is in the form of English language. So, different techniques are developed for management of English documents. But due to the development of regional languages the data in regional languages is also available in valuable amount. The vast amount of data causes problems in its management. So, Automatic classification of documents raises much more attention in few decades. Automatic classification is the task of assigning predefined category to the unlabelled documents. This gives the class to the document to which it actually belongs. Automatic classification frees the organizations to handle the large amount of documents manually and enhance the retrieval process. It can be concluded that number of classification techniques are available. But for different data length same technique gives different accuracy.

***Index Terms*—Automatic Classification, K nearest Neighbour (KNN), Naive Bayes (NB), Neural Network, Vector Space Model (VSM).**

I. INTRODUCTION

Now-a-days, classification is the major task to handle the vast amount of data in proper manner. Automatic classification is the process of assigning predefined categories to the uncategorized documents. Automatic classification assigns the label to unlabelled documents. After label assignment they are considered to the particular category. Automatic Classification increases the readability of documents by decreasing its retrieval time. Automatic classification frees the organizations from manually work which are very time consuming process [1], [2]. This paper describes the different uses of classification, issues in the classification and various classification techniques.

II. APPLICATIONS OF DOCUMENT CLASSIFICATION

A. *Document classification is need of every organization and it has following applications [3], [4]:*

1. ***Fast Retrieval*:** Document classification enables the fast retrieval of documents related to particular category. From uncategorized documents, it is very difficult to find the particular document. Automatic classification increases the rate of retrieval of documents of particular category.
2. ***Sentimental Analysis*:** Document classification also determines the attitude of writer or speaker of particular document. Automatic classification helps for sentimental analysis, which gives the reaction of reader with respect to the documents of various categories.
3. ***Knowledge Reuse:*** Document classification enhances the knowledge reuse. The user that wants the knowledge about the various topics does not suffer with the problems of finding them from unstructured data. Knowledge of one document of specific category is reused efficiently and effectively by automatic classification.
4. ***Readability Assessment:*** Document classification maximizes the readability of documents. Automatic classification is the task that categorizes the documents, so it maximizes the readability of documents of particular category.
5. ***Automated Indexing:*** Documents are automatically indexed when they categorized into particular category.

III. ISSUES IN DOCUMENT CLASSIFICATION

Document classification is necessary but at same time there are many issues in this field. Document classification has following issues [5]:

1. ***Document Type*:** Document type refers to the content in document and the size of document. The documents of same category may give variations in results depends upon the size of them and the language used in them.

2. ***Classifier Type:*** When documents are classified the type of classifier used is necessary question. For different types of documents with different length, the different types of classifier are used.

IV. CLASSIFICATION MODEL

To design the classifier the basic needs are to collect the data and labels. Then find the features from these data sets and choose any classification technique. Train the classifier with collected data and then test the accuracy of classifier with test data which is odd one from the train data [6], [7]. Classification Model is represented by Fig. 1.

Classification model consist of two phases training phase and testing phase.

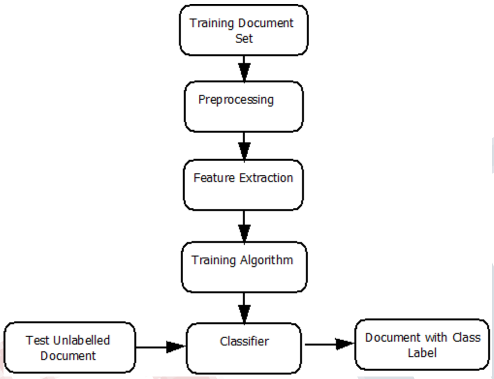

Fig. 1 Classification Model

1. ***Training phase***: Training phase consist of set of documents which consists the predefined labels on the basis of their features such as terms and content. Documents in training set are labeled documents which help to classify the new documents.
2. ***Testing phase***: Testing phase consist the set of documents which are initially unlabelled and given to the trained system to assigning them labels based on their terms and content.

These two phases further contains the three sub phases:

i. ***Pre-processing phase:*** This phase includes tokenization, digit removal, punctuation removal, stop words removal and stemming.
ii. ***Feature extraction phase***: It consist of statistical approach to extract relevant features from the documents.
iii. ***Processing steps:*** This last phase applies text classification algorithms to the extracted features to classify the documents into classes in testing phase.

V. CLASSIFICATION TECHNIQUES

Classification is major need to handle the different types of documents. So there are many classification techniques are developed.

- ❖ Rule Based Classification
- ❖ K Nearest Neighbour
- ❖ Naive Bayes
- ❖ Neural Networks
- ❖ Support Vector Machine
- ❖ Semantic Based Classification
- ❖ Decision Tree
- ❖ Vector Space Model

Rule Based Classification: Rule based classification is the simplest technique amongst all techniques. In this the classification is performed on the basis of formation of rules. The train set is used for the formation of rules and then those rules are used for test set. The rules are formed by different methods based on the categories. Rules in rule based classification stores in knowledge base and those rules are retrieved when test documents are given to the classifier. The classifier gives the results according to the rules in knowledge base [4].

K Nearest Neighbour: KNN is instance learning technique. In this classification technique the classification is based on the k set of samples. In KNN predefined samples are used for training purpose. The new sample is categorized on the bases of existing training samples by measuring the distance between them without considering sample labels. The label is assigned to the new sample similar to those with which its distance measure is smaller. In KNN train set contains the sample with features such as x_1, x_2, x_3… x_k and the new sample y is categorized on the basis of train set samples as f(y) = { x_1, x_2, x_3… x_k }.

KNN generally works on the distance metrics. The most common distance metric is used Euclidean distance as in equation (1), other variations in Euclidean distance are Manhattan and Minkowski distance measures shown by equations (2) and (3) respectively. Large the value of k more precise results are developed [8]-[11].

$$\text{Euclidean Distance: } \sqrt{\sum_1^k (x_i - y_i)^2} \qquad (1)$$

$$\text{Manhattan: } \sum_1^k |x_i - y_i| \qquad (2)$$

$$\text{Minkowski: } \left(\sum_1^k (|x_i - y_i|)^q\right)^{\frac{1}{q}} \qquad (3)$$

Naive Bayes: Naive Bayes is supervised learning technique based on Bayes theorem with naive assumption of independence between the set of pairs of features. Naive Bayes classification is based on the posterior probability and maximum likelihood calculations of the documents as shown by equations (4) and (5).

$$p(c_k|x) = \frac{p(c_k)*p(x|c_k)}{p(x)} \qquad (4)$$

$$posterior = \frac{prior*liklihood}{evidence} \qquad (5)$$

In this the documents are treated as bag of words and then for each word, respective posterior probability and likelihood is calculated. Then the Naive Bayes theorem is applied to find the class label of particular document. Naive Bayes is generally works on the feature extracted from the document. The feature assumption made the features irrelevant; consequently one feature does not affect another feature. The Naive Bayes classifier is trained by small amount of train data to estimate features required for classification process [10], [16]-[18].

Neural Networks: Neural network is set of interconnected nodes called neurons which are inspired from human neuron system. Neural network is the layered web which contains set of nodes and layers. A neural network contains the set of layers such as input layer, hidden layer and output layer. Input layer is used for input and a hidden layer performs the complex computations which are necessary for the classification and output layer gives the classification results.

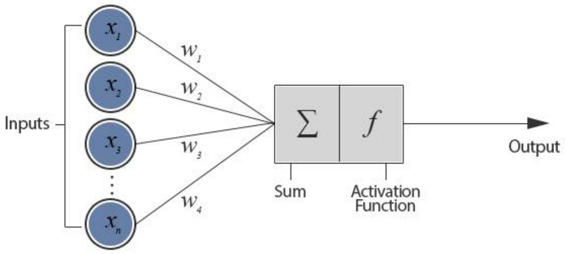

Fig. 2 Neural Networks

Nodes in the neural network are connected by the set of weighted vectors. Each node contains the different activation scores. Activation scores of a node are the combination of incoming vector weights. Different nodes have different activation scores which are used for classification purpose [12]-[15]. Neural Networks are represented by Fig. 2.

***Support vector Machine*:** Support vector machine is supervised classification technique, in which the documents are classified on the basis of hyperplanes. In this classification technique training data set is given with labels and algorithm outputs the hyperplanes which classify the test data set into predefined class labels. In support vector machine, those hyperplanes are considered which passes through the maximum train data set. It is the vector space based technique, in which the goal is to be find decision boundary (hyperplane) between two documents to assign class label to them [19]-[22].

***Semantic Based Classification*:** Semantic Based classification techniques use the semantic relationship between the words of documents. In this method the documents are treated as the bag of words and then classification is performed on the basis of relationship between them. Semantic based techniques use the word net approach for classification. Different wordnet approaches are used for classification technique. Semantic based classification considers the relationship between the words of documents and also depends upon the meaning of each word. In this documents are compared by the relationships and the meanings of words. Similarity between two documents is based on the meaning of the words included in them. If the words of different documents are shows the similar meanings then they are classified as of same class label [25]-[27].

Decision Tree: Decision tree is the simplest classification technique. In this, training documents are manually classified by using simple true/false queries. Decision tree consist of nodes (root node and leaf nodes) and links between those nodes. The leaf node of tree defines the class label and links defines features that lead to those class labels. In decision tree classification the documents are given to the root node and goes down to the leaf nodes by satisfying the queries on the links. At the end document goes to the leaf node which represent its class label most appropriately [6], [23], [24]. Decision tree is represented by Fig. 3.

The key requirements of decision tree are as following.

- ❖ Attribute value description
- ❖ Predefined classes
- ❖ Sufficient data

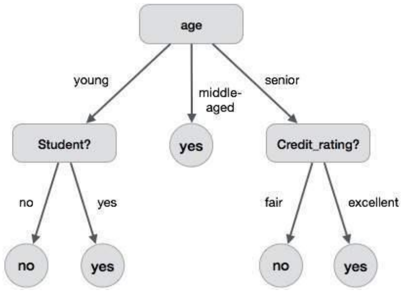

Fig. 3 Decision Tree

***Vector Space Model*:** Vector space model is an algebraic model in which the text documents are represented as vectors. Vector space model is used for information filtering, retrieval and classification. In vector space model the term of documents are represented as axis in space and the documents are represented as vector of the sum of all terms in that document. Vector space model use the train set and test set to classify the documents. In this, term frequency of each term in the documents is calculated. Term frequency is number of times the term appears in the document [28].

Vector space model also find the inverse document frequency of each term in the document. The inverse document frequency shows that the rare terms i.e. terms with smaller term frequency also important to the document. Then by using term frequency (TF) and inverse document frequency (IDF) weight is assigned to each term. Vector space model use various coefficients to find out similarity. One of these coefficients is selected according to the problem to assign class label to the input document on the basis of calculated weights. In this weights are not binary and calculations are very easy to compute [11], [12], [18], [29], [30]. Vector Space Model is represented by Fig. 4.

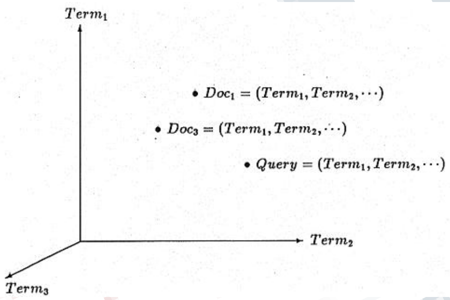
$Term_1$
$Doc_1 = (Term_1, Term_2, \cdots)$
$Doc_3 = (Term_1, Term_2, \cdots)$
$Query = (Term_1, Term_2, \cdots)$
$Term_2$
$Term_3$

Fig. 4 Vector Space Model

VI. CONCLUSION

Document classification is the one of necessary task for fast retrieval of documents of particular category. There are various classification techniques, but which technique best fits to the input data depends upon document length. For large data set K Nearest Neighbour and Support Vector Machine classification techniques are useful. K Nearest Neighbour is useful for testing in large data sets while Support Vector Machine is useful for training purposes. For reasonable size of data set again Support Vector Machine gives the best results. Naive Bayes and Decision Tree are useful for small data sets.

REFERENCES

[1] S. Fabrizio, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[2] N. Lyfenko, "Automatic classification of documents in a natural language: A conceptual model," *Automatic Documentation and Mathematical Linguistics*, vol. 48, no. 3, pp. 158–166, 2014.

[3] A. Chopra, A. Prashar, and C. Sain, "Natural language processing," *International Journal of Technology Enhancements and Emerging Engineering Research*, vol. 1, no. 4, pp. 131–134, 2013.

[4] Nidhi and V. Gupta, "Recent trends in text classification techniques," *International Journal of Computer Applica- tions (0975 8887) Volume*, vol. 35, 2014.

[5] H. S. Christopher D. Manning, Prabhakar Raghavan, *An Introduction to Information Retrieval*, 2009.

[6] C. C. Aggarwal, *Data Classification Algorithms and Applications*, 2015.

[7] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of advances in information technology*, vol. 1, no. 1, pp. 4–20, 2010.

[8] S. Tan, "Neighbor-weighted k-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, no. 4, pp. 667–671, 2005.

[9] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[10] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[11] W. M. HADI, T. FADI, AND H. ABDEL-JABER, "a comparative study using vector space model with k-nearest neighbor on text categorization data." in *world congress on engi- neering*, 2007, pp. 296–300.

[12] k. rajan, v. ramalingam, m. ganesan, s. palanivel, and b. palaniappan, "automatic classification of tamil documents using vector space model and artificial neural network," *expert systems with applications*, vol. 36, no. 8, pp. 10 914–10 918, 2009.

[13] f. harrag and e. al-qawasmah, "improving arabic text categorization using neural network with svd." *jdim*, vol. 8, no. 4, pp. 233–239, 2010.

[14] t. d. sanger, "optimal unsupervised learning in a single layer linear feedforward neural network," *neural networks*, vol. 2, no. 6, pp. 459–473, 1989.

[15] l. manevitz and m. yousef, "one-class document classification via neural networks," *neurocomputing*, vol. 70, no. 7, pp. 1466–1481, 2007.

[16] f. jensen, "an introduction to bayesian networks springer," *new york*, 1996.

[17] j. chen, h. huang, s. tian, and y. qu, "feature selection for text classification with na¨ive bayes," *expert systems with applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[18] d. isa, l. l. hong, v. kallimani, and r. rajkumar, "text document pre-processing using the bayes formula for classification based on the vector space model," *computer and information science*, vol. 1, no. 4, p. 79, 2008.

[19] t. joachims, "transductive inference for text classifica- tion using support vector machines," in *icml*, vol. 99, 1999, pp. 200–209.

[20] a. sun, e. p. lim, and y. liu, "on strategies for imbalanced text classification using svm: a comparative study," *decision support systems*, vol. 48, no. 1, pp. 191– 201, 2009.

[21] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45– 66, 2002.

[22] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *Neural Networks, IEEE Transactions on*, vol. 10, no. 5, pp. 1048–1054, 1999.

[23] a. chidanand and d. fred, "automated learning of decision rules for text categorization," *acm trans. inf. syst.*, vol. 12, no. 3, pp. 233–251, jul. 1994.

[24] M. N. Anyanwu and S. G. Shiva, "Comparative analysis of serial decision tree classification algorithms," *Interna- tional Journal of Computer Science and Security*, vol. 3, no. 3, pp. 230–240, 2009.

[25] R. mihalcea, c. corley, and c. strapparava, "corpus- based and knowledge-based measures of text semantic similarity," in *aaai*, vol. 6, 2006, pp. 775–780.

[26] s. parseh and a. baraani, "improving persian document classification using semantic relations between words," *arxiv preprint arxiv:1412.8147*, 2014.

[27] m. á. corella and p. castells, "*semi-automatic semantic-based web service classification*," in *business process management workshops*. springer, 2006, pp. 459–470.

[28] A. Aizawa, "An information-theoretic perspective of tf– idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[29] G. Salton, A. Wong and C.S. Yang, "A vector space model of automatic indexing", *Communication of the ACM*, vol. 18, no. 11, pp. 615-620, 1975.

[30] P. Gawande and P. A. Suryawanshi, "Improving Web Page Classification by Vector Space Model," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2015, Apr. 2015. [Online].Available:http://www.rroij.com/abstract.php ?abstract id=44783.