Journal of Advances in Mathematics and Computer Science

29(6): 1-11, 2018; Article no.JAMCS.34128
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Web Document Classification Using Naïve Bayes

A. B. Adetunji[1], J. P. Oguntoye[1*], O. D. Fenwa[1] and N. O. Akande[2]

[1]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Ladoke Akintola University of Technology (LAUTECH), Nigeria.
[2]Department of Physical Sciences, College of Science and Engineering, Landmark University Omu-Aran, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Author ABA and ODF designed the study, wrote the protocol and supervised the work. Authors JPO and NOA carried out all laboratories work and performed the statistical analysis, Author JPO wrote the first draft of the manuscript. Author ABA and JPO managed the analyses of the study. Authors ODF and NOA managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2018/34128
Editor(s):
(1) Dr. Kai-Long Hsiao, Associate Professor, Taiwan Shoufu University, Taiwan.
(2) Dr. Tian-Xiao He, Professor, Department of Mathematics and Computer Science, Illinois Wesleyan University, USA.
Reviewers:
(1) Ismail Olaniyi Muraina, Adeniran Ogunsanya College of Education, Nigeria.
(2) R. Nedunchelian, Anna University, India.
(3) S. Sridhar, R.V.College of Engineering, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/27781

Original Research Article

Received: 15 May 2017
Accepted: 22 August 2017
Published: 16 December 2018

Abstract

World Wide Web has become a huge collection of documents and the amount of documents available is increasing on a daily basis. How to correctly classify the vast documents into a particular category and locate any document of interest easily has become a challenge researchers have been trying to solve for decades and different researchers have attempted different algorithms using different platform to achieve this aim. In this paper, a University web site was used as a case study and a machine learning workbench called WEKA (Waikato Environment for Knowledge Analysis) which provides a general-purpose environment for automatic classification, regression, clustering and feature selection was used as a machine learning platform. Running Naïve Bayes with 10-fold cross validation on the selected web data gives a 77% correctly classified instances in zero second with relative absolute error of 68.9937%. This shows the ability of Naïve Bayes algorithm to accurately classify vast amount of web document in a short time.

**Corresponding author: E-mail: jonatoye2008@gmail.com;*

Keywords: Machine learning; Naïve Bayes; web document classification; Waikato Environment for Knowledge Analysis (WEKA).

1 Introduction

The Internet is a vast resource of information of different types: text, images, audio and video [1]. The amount of information available on the World Wide Web (WWW) has been increasing at an exponential rate. These 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.

According to Pierre [2], the number of web pages available on the web is around 1 billion and another 1.5 million are being added on a daily basis, this explosive growth rate has put huge amounts of information at the disposal of anyone with access to the Internet. Hence, how to access a particular web document out of these enormous web pages available on the internet and how to correctly classify them has being a problem researchers have been trying to solve. Though different search engines are available, they do not provide the exact information that matches at a high degree of relevance what the user's interests and preferences are simply because the information available on the internet are not well organized. This has led to a great deal of interest in developing useful and efficient tools that can be used to properly organize web pages.

The World Wide Web continues to grow both in the huge volume of traffic and the size and complexity of Web sites. It is difficult to identify the relevant information present in the web [3]. With the growing number of web documents and online information, web mining plays an important role in extracting useful information from the World Wide Web through Web page classification, also known as web page categorization. Web page classification may be defined as the task of determining whether a web page belongs to a particular category or not.

Web content mining is nothing but the discovery of valuable information from web documents and these web documents may contain text, image, hyperlinks, metadata and structured records [4]. Web mining is applied to extract the interesting, useful patterns and hidden information from the Web documents and Web activities [5].

In this paper, a method of automatically classifying Web documents into a set of categories using the Naive Bayes algorithm is proposed and Waikato Environment for Knowledge Analysis (WEKA) is used as the machine learning platform. The outline of this paper is the following. In section 2 we review related works and in section 3, we introduced our classification algorithm which is the Naïve Bayes. Our research methodology was described in detail in section 3 while our research methodology was discussed in the following section. Our result was discussed in section 5. We concluded our paper in the last section.

2 Review of Related Works

The Web mining can be said to have three operations of interests: Clustering (e.g., finding natural grouping of users, pages, etc.), Association (e.g., which URLs tend to be requested together), Sequential Analysis (e.g., the order in which URLs tends to be accessed). The clusters and associations in web mining do not have clear-cut boundaries and often overlap considerably in most real world problems [6].

Consequently, an increasing number of approaches have been developed for web document classification, including k-nearest-neighbour (KNN) classification [7,8,9], Naïve Bayes classification [10,11,12], Support Vector Machines (SVM) [13,14,15], decision tree (DT) [16,17], Neural Network (NN) [18,19] and maximum entropy [8,20].

Among these approaches, the Naïve Bayes text classifier has been widely used because of its simplicity in both the training and classifying stage [14]. The naive Bayesian classifier is uncomplicated and widely used method for supervised learning. It is one of the fastest learning algorithms, and can deal with any number of

features and classes [21]. Bayesian classification is based on Bayes theorem. A simple Bayesian classification namely the Naïve classifier is comparable in performance with decision tree and neural network classifiers [22].

Loan [23] submitted that Naïve Bayes algorithm improves the tasks of the Web Mining by its accurate classification of web documents. Its applications are important in the following areas: e-mail spamming; filtering spam results out of search queries; mining log files for computing system management; machine learning for Semantic Web; document ranking by text classification; hierarchical text categorization; managing content with automatic classification and other areas from Web Mining.

Ziqiang and Xia [24] proposes a web classification algorithm using Maximum Margin Projection (MMP) and Least Square Support Vector Machines (LS-SVM). The high-dimensional document data is first projected into lower-dimensional feature space via MMP algorithm, then, the LS-SVM classifier is used to classify the test documents into different class in terms of the extracted semantic features. Experiments performed on two popular document datasets demonstrate the superior performance of the proposed document classification algorithm.

Guan, Zhou, Xiao, Guo and Yang [25] introduced a Fast dimension reduction for document classification based on imprecise spectrum analysis. It uses a representative matrix composed of top-k column vectors derived from the original feature vector space and reduces the dimension of a feature vector by computing its product with this representative matrix. Howard, Paull, Biletskiy and Yang [19] developed a fast back-propagation neural network model to build document classifiers and the information gain method is used for feature selection. According to the rank of the information gain of all the words contained in the documents, those words that contain more information to classify the documents were selected as the input features of the artificial neural network (ANN) classifiers. The neural network developed assumes a three-layer structure with a fast back-propagation learning algorithm.

Rujiang and Xiaoyue [26] proposed a system that uses integrated ontologies and natural language processing techniques to index texts. The traditional words matrix is replaced by a concepts-based matrix. For this purpose, a fully automated method for mapping keywords to their corresponding ontology concepts was developed using SVM for classification. Their results show an improved text classification performance. In this paper, we propose a novel approach to classifying web document using Naïve Bayes10-fold cross validation. The data used was extracted from the Website of Ladoke Akintola University of Technology (LAUTECH), Ogbomoso, Nigeria. WEKA was used as the machine learning workbench which provides a general-purpose environment for automatic classification and feature selection.

3 Naive Bayes Approach

Naive Bayes is the simplest Bayesian Network (BN) Classifier, in which each attribute node (which is the attribute variable) has the class node (which is the class variable) as its parent, but does not have any other parent.

The Naïve Bayes classifier applies to learning tasks where each instance *x* is described by a conjunction of attribute values and where the target function *f(x)* can take on any value from some finite set V. A set of training examples of the target function is provided and a new instance is presented, described by the tuple of attribute values $\{a_1, a_2, \ldots, a_n\}$ this will predict the target value, or classification, for this new instance. Using Bayesian approach in classifying the new instance means assigning the most probable target value, V_{MAP} given the attribute values $\{a_1, a_2, \ldots, a_n\}$ that describes the instance.

$$V_{MAP} = \underset{v_i \in V}{\operatorname{argmax}} P(v_i | a_1, a_2, \ldots, a_n) \quad \text{where} \quad v_i \in V \tag{1}$$

The Bayes theorem can be used to rewrite the expression above as

$$V_{MAP} = \underset{v_i \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, \ldots, a_n | v_i) P(v_i)}{P(a_1, a_2, \ldots, a_n)} \quad (2)$$

$$V_{MAP} = \underset{v_i \in V}{\operatorname{argmax}} P(a_1, a_2, \ldots, a_n | v_i) P(v_i) \quad (3)$$

Now we could attempt to estimate the two terms in equation (3) based on the training data. It is easy to estimate each of the $P(v_i)$ simply by counting the frequency with which each target value v_i occurs in the training data. The Naïve Bayes classifier is established on the basic postulation that the characteristic values are conditionally independent with respect to a target value. In other words, the assumption is that given the target value of the instance, the probability of observing the conjunction $a_1, a_2, \ldots, a_n$ is just the product of the probabilities for the individual attributes:

$$P(a_1, a_2, \ldots, a_n | v_i) = \prod_i P(a_1 | v_i) \quad (4)$$

Substituting this into equation (3) we have Naïve Bayes Classifier:

$$V_{NB} = \underset{v_i}{\operatorname{argmax}} P(v_i) \prod_{i=1}^{N} P(a_1 | v_i) \quad (5)$$

Where V_{NB} denotes the target value output by the Naïve Bayes Classifier.

4 Waikato Environment for Knowledge Analysis (WEKA)

The Waikato Environment for Knowledge Analysis (WEKA) is a machine learning workbench currently being developed at the university of Waikato. Its purpose is to allow users to access a variety of machine learning techniques for the purposes of experimentation and comparison using real world data sets. Weka is a comprehensive suite of java class libraries that implement many state-of-the-art machine learning and data mining algorithms. WEKA is freely available on the world-wide web and accompanies a new text on data mining which documents and fully explains all the algorithms it contains [27]. Applications written using the WEKA class libraries can be run on any computer with a web browsing capability; this allows users to apply machine learning techniques to their own data regardless of computer platform.

5 Research Methodology

The research methodology for this study involves the following steps which include Data collection, Data preparation and the Machine Learning. This section summarizes these steps.

5.1 Data collection

Generally, the data collection step involves gathering text or web documents. The web document used in this study were collected from LAUTECH website. Fig. 1 illustrates a web page sample from of the data collected. The web page consists of texts, hyperlinks and pictures. Consequently, data pre-preparation will be needed to remove other element of the web page order than text. This is necessary due to the fact that the HTML structure sometimes have semantics associated with the document class. Therefore, the HTML structure is ignored so as to simplify the document processing and document representation. Also, the web page used were classified into categories and a class label.

5.2 Data preparation stage

After extracting the texts in the web pages, the data collected was converted into data sets in Weka's Attribute Relation File Format (ARFF) to be later used in the Machine Learning phase.

All text or web documents (text corpus) obtained from the data collection step were concatenated and saved in a single text file where each document is represented on a separate line in plain text format. This representation uses three attributes: document_name, document_content, and document_class, all of type string.

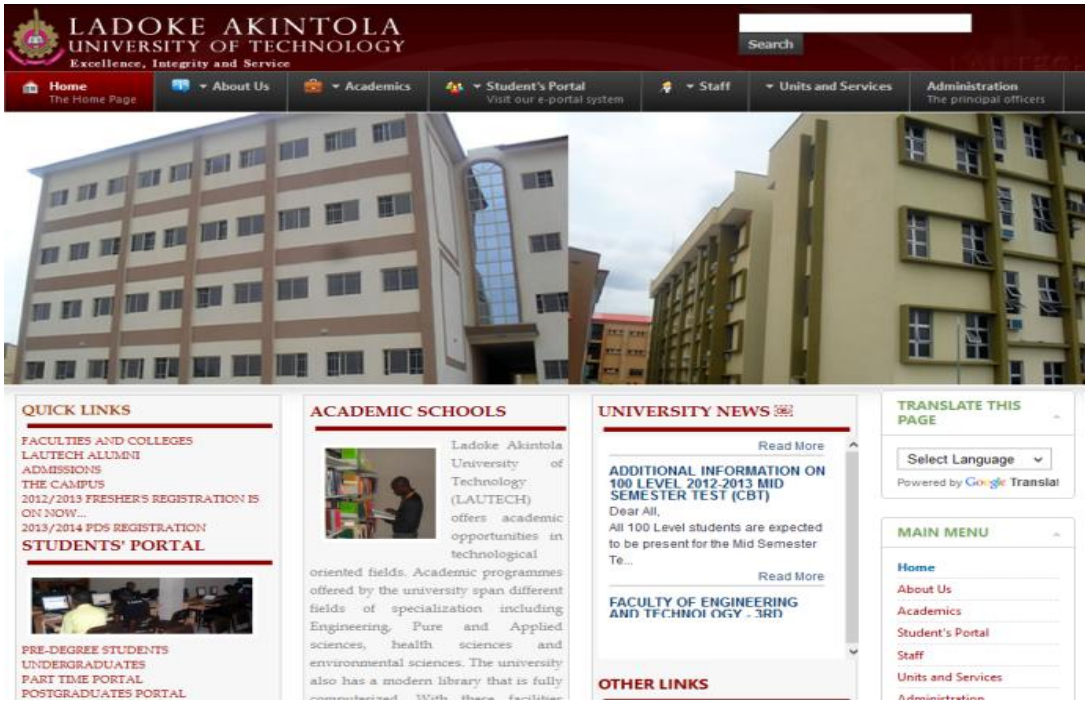

Fig. 1. Web page sample

5.2.1 Data convertion into .arff format

To use WEKA as the machine learning tool, the data to be used must be in .arff format. WEKA provides three ways of data conversion which are: Excel, Notepad and Ms Word. In this study, the collected data was converted to.arff format using the notepad.

5.2.2 Data classification

In this study, two types of data were collected, the first set of data are a set of pages describing the different units that exist in LAUTECH such as ICT Centre, physical and planning units, academic planning unit, the health centre, sports development unit, works and maintenance unit.

The second sets of data are a set of data describing some of the departments we have in LAUTECH such as biology, physics, computer, chemistry, fine arts, general studies, accounting, earth science etc. Hence, data are in two categories.

5.3 Machine learning phase

WEKA is the machine learning tools used in this study for web document classification. The prepared data in .arff format was loaded into WEKA and Naïve Bayes algorithm was applied to the data i.e. after the data had been converted from string to nominal form. Classifying text document, the attribute content is usually

much higher than the attribute name, this leads to the problem of having too many 0's in the document-term matrix, hence, a subset of words (bag of Words) that best represent the document collection with respect to the classification task was created. The process of removing these unwanted elements is called Feature (attribute) Selection. Weka provides a good number of algorithms for this purpose which is available through the attribute selection filter.

Table 1. Data classification

Departmental web pages	Classes (A-Units, B- Departments)
ICT Centre, Physical and Planning Units, Academic Planning Unit, The Health Centre, Sports Development Unit, Works and Maintenance Unit	Class A
Biology, Physics, Computer Science, Chemistry, Fine Arts, General Studies, Earth Science, Math	Class B

Three file header which are: document_name, document_content and document_class were used. The document class has two classes i.e. class A and B displayed in two different colours which are red and blue. The classes in blue colour belong to class A while those in red colours belong to class B. The order of the arrangements shows the order in which the web pages occur on the LAUTECH Website.

6 Results and Discussion

The network structure describes the structure of the data used. Each of the variables is followed by a list of parents, so the class variable has parent document_class, the number in braces is the cardinality of the variable. It shows that in the dataset there are three class variables. All other variables are made binary by running it through a discretization filter

6.1 Log result

The logarithmic score shows the logarithmic values of the network structure for various methods of scoring.

Table 2. Log result table

Log	Values
LogScore Bayes	-98.17916871173622
LogScore BDeu	-2811.984944106161
LogScore MDL	-1337.403671923017
LogScore ENTROPY	-729.4464017755708
LogScore AIC	-1178.4464017755708

Time taken to build model: 0 seconds

6.2 Stratified cross-validation

The stratified cross-validation shows 77% correctly classified instances, 23% incorrect classified instances, 0 kappa statistic, 0.4838 mean absolute error, 0.3108 root mean square error, 68.9937% relative absolute error and 100 total number of instances.

6.3 Detailed accuracy by class

From Table 4 two important observations can be made; First, all attributes have only one of its values occurring in class B. This is indicated by the fact that one of the counts is always 1, which means that the actual count is 0 (according to the Laplace estimator used by the algorithm the actual value count is incremented by 1).

Table 3. Log result table

Items	Number	Percentage
Correctly Classified Instances		77%
Incorrectly Classified Instances		23%
Kappa statistic	0	
Mean absolute error	0.4838	
Root mean squared error	0.3108	
Relative absolute error		68.9937%
Total Number of Instances	100	

Table 4. Detailed accuracy by class table

Class	TP Rate	FP Rate	Precision	Recall	F- Measure	ROC Area
A	0.833	0.353	0.821	0.833	0.827	0.82
B	0.647	0.167	0.667	0.647	0.657	0.82
Weighted Avg.	0.770	0.290	0.768	0.770	0.769	0.82

=== Confusion Matrix ===

a b <-- classified as

55 11 | a = c0

12 22 | b = c1

Second, the confusion matrix indicates that 55 of 66 instances in class A were accurately classified while 22 of 34 instances in class B were accurately classified. This shows the ability of WEKA to correctly classify documents using Naïve Bayes classifier.

6.4 Predictions on test data

In Table 5; for all documents from actual class B, the class distribution decisively predicts class B that means all the documents under class B were accurately classified but the predictions show that the six errors (marked with +) happen in actual class A.

Table 5. Predictions on test data

Inst#	Actual	Predicted	Error	Probability distribution
1	1:A	2:B	+0.469	*0.531
2	2:B	2:B	0.469	*0.531
1	1:A	2:B	+0.469	*0.531
2	2:B	2:B	0.469	*0.531
1	1:A	2:B	+0.469	*0.531
2	2:B	2:B	0.469	*0.531
1	1:A	2:B	+0.469	*0.531
2	2:B	2:B	0.469	*0.531
1	1:A	2:B	+0.469	*0.531
2	2:B	2:B	0.469	*0.531
1	1:A	2:B	+0.464	*0536
1	2:B	2:B	0.483	*0.517
1	2:B	2:B	0.483	*0.517
1	2:B	2:B	0.483	*0.517
1	2:B	2:B	0.483	*0.517

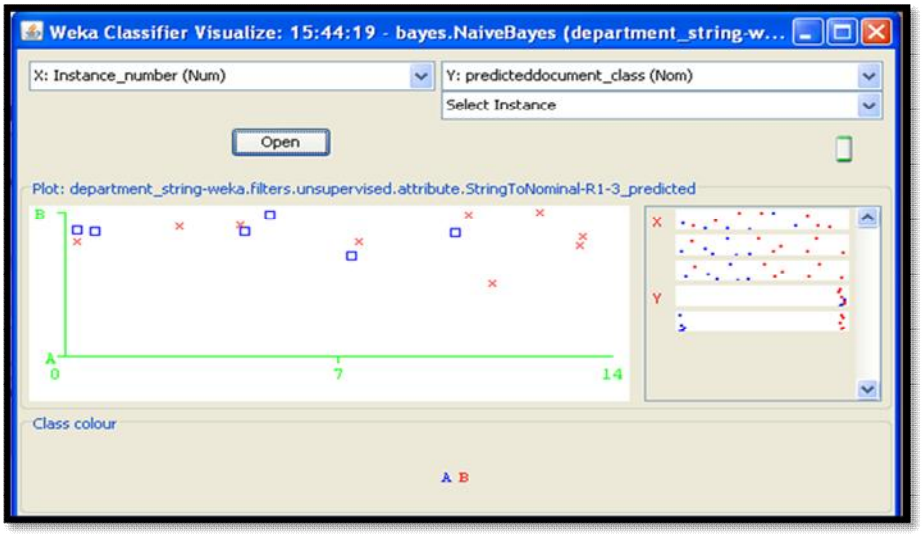

Fig. 2. Graphical representation of form showing the six error of the classifier

Clicking on the first two squares in the plot reveals these two documents as shown in Fig. 3 below:

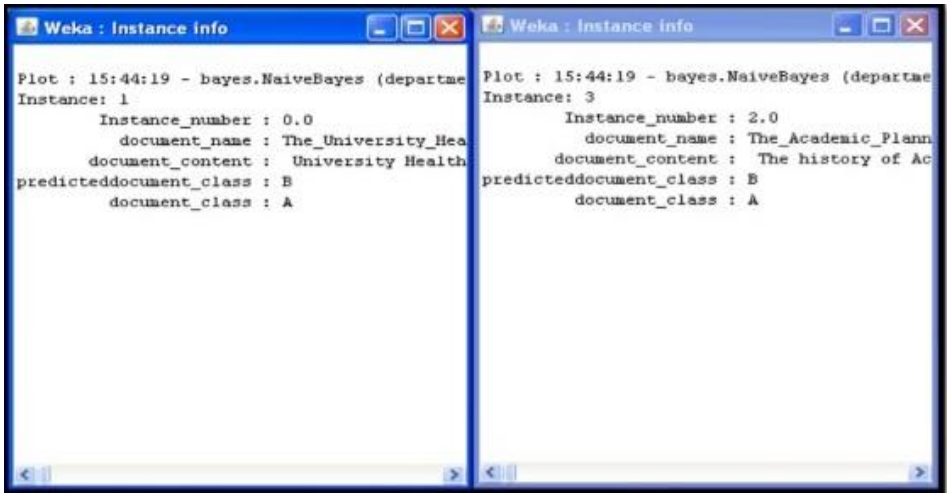

Fig. 3. Form showing the instance information of the wrongly classified documents

These forms show the first and second documents that were classified wrongly, both belongs to class A but were wrongly classified as class B.

Table 6. Comparison of Naïve Bayes accuracy with the existing classifiers

Author	Method	Accuracy (%)
Materna J. [28]	Support Vector Machine	79.04
Kavitha et al. [29]	Artificial Bee Colony (ABC)	84.00
Mahmoud et al. [30]	NB & SVM & KNN	99.98
Markov et al. [31]	Naïve Bayes classifier	77.00
Developed	Naïve Bayes classifier	77.00

7 Conclusion

In this Study, the strength of Naïve Bayes classifier in classifying web documents was discovered and WEKA by the virtue of its performance in classifying web document is a good machine learning environment for web mining. The main strength of this approach lies in its ability to correctly classify the web documents into the right categories and its ability to classify web pages in a short time of zero seconds. The result obtained can be improved to achieved an increased accuracy of a web page classification by combining other techniques like Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN).

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Aldekhail M. Application and significance of web usage mining in the 21st century: A literature review. International Journal of Computer Theory and Engineering. 2016;8(1).
DOI: 10.7763/IJCTE.2016.v8.1017

[2] Pierre JM. Practical issues for automated categorization of web pages. Proceedings of the 4th international workshop on Web information and data management. McLean, Virginia, USA. 2000; 96–99.

[3] Jayalatchumy D, Thambidurai P. Web mining research issues and future directions-a survey. IOSR Journal of Computer Engineering. 2013;14(3):20-27.

[4] Vijiyarani S, Suyanya E. Research issue in web mining. International Journal of Computer-Aided Technologies. 2015;2(3).
DIO: 10.5121/ijcax.2015.2305

[5] Arti Choudhary S, Purohit GN. Role of web mining in e-commerce. International Journal of Advanced Research in Computer and Communication Engineering. 2015;4(1).
DOI: 10.17148/IJARCCE.2015.4155

[6] Kumar SN. World towards advance web mining: A review. American Journal of Systems and Software. 2013;2(3):44-61.

[7] Han EH, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbour classification. Department of Computer Science and Engineering, Army HPC Research Center. University of Minnesota; 1999.

[8] Kwon OW, Lee JH. Text categorization based on k-nearest neighbour approach for web site classification. Information Processing and Management. 2003;29(1):25-44.

[9] Calado P, Cristo M, Moura E, Ziviani N, Ribeiro-Neto B, Goncalves MA. Combining link-based and content-based methods for web document classification. In CIKM Proceedings of the 12th International Conference on Information and Knowledge Management, New York. 2003;394-401. ACM Press.

[10] McCallum A, Nigam K. A Comparison of event models for naïve bayes text classification. Journal of Machine Learning Research. 2003;3:1265-1287.

[11] Denoyer L, Gallinari P. Bayesian network model for semi-structured document classification. Information Processing and Management. 2004;40:807–827.

[12] Wang L, Ji P, Jing Q, Siqing S, Zhuming B, Weiguo D, Naijing Z. Feature weighted naïve Bayes algorithm for information retrieval of enterprise systems. Enterprise Information Systems. 2014;8(1): 107-120.

[13] Rung-Ching C, Chung-Hsun H. Web page classification based on a support vector machine using a weighted vote schema. Expert Systems with Applications. 2007;31:427–435.

[14] Sun A, Lim E, Ng W. Web classification using support vector machine; 2002.

[15] Zhang D, Lee WS. Web taxonomy integration using support vector machines. WWW '04:Proc. of Int. Conf. on World Wide Web. 2004;472-481. ACM Press.

[16] Estruch V, Ferri C, Hern´andez-Orallo J, Ramırez-Quintana MJ. Web categorisation using distance-based decision trees. Electronic Notes in Theoretical Computer Science. 2006;157:35–40.

[17] Tian Y, Huang T, Gao W, Cheng J, Kang P. Two-phase web site classification based on hidden markov tree models. In WI ’03: Proceedings of the IEEE/WIC International Conference on Web Intelligence, Washington, DC, USA. 2003;227. IEEE Computer Society.

[18] Manevitza L, Yousef M. One-class document classification via Neural Networks. Neurocomputing. 2007;70:1466–148.

[19] Howard L, Paull L, Biletskiy Y, Yang S. Document classification using information theory and a fast back-propagation neural network. Intelligent Automation AND Soft Computing. 2010;16(1):25-38.

[20] Chieu HL, Ng HT. Maximum entropy approach to information extraction from semi-structured and free text. Proceedings of the Eighteenth National Conference on Artificial Intelligence. 2002;786-791.

[21] Mahesh KM, Saroja DH, Prashant GD, Niranjah C. Text mining approach to classify technical research documents using naïve bayes. International Journal of Advanced Research in Computer and Communication Engineering. 2015;4(7).
DOI: 10.17148/IJARCCE.2015.4789

[22] Anuradha P, Deepika A, Payal J, Priyanshi A. Text classification in data mining. International Journal of Scientific and Research Publications. 2015;5(6).

[23] Loan P. An approach of the naive bayes classifier for the document classification. General Mathematics. 2006;14(4):135–138.

[24] Ziqiang W, Xia S. Document classification algorithm based on MMP and LSSVM. Procedia Engineering. 2011;15:1565–1569.

[25] Guan H, Zhou J, Xiao B, Guo M, Yang T. Fast dimension reduction for document classification based on imprecise spectrum analysis. Information Sciences. 2013;222:147–162.

[26] Rujiang B, Xiaoyue W. Using an integrated ontology database to categorize web pages. Journal of the Chinese Institute of Engineers. 2012;35(5):509-514.

[27] Witten IH, Frank E. Data mining: Practical machine leaming; tools and techniques with java implementations. Morgan Kaufmann, San Francisco; 1999.

[28] Materna J. Automatic web page classification. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. 2008;84–93.

[29] Kavitha C, Sudha Sadasivam G, Kiruthika S. Semantic similarity based web document classification using Artificial Bee Colony (ABC) algorithm. WSEAS Transactions on Computers. 2014;13:476-484.

[30] Mahmoud TM, Abd-El-Hafeez T, Nour El-Deen DT. A design of an automatic web page classification system. British Journal of Applied Science & Technology. 2016;18(6):1-14.

[31] Markov A, Last M, Kandel A. Model-based classification of web documents represented by graphs. WEBKDD'06 Philadelphia, Pennsylvania, USA; 2006.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/27781