# Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (1): 10-17, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.10.17


# Bayesian Folder Allocation System for Electronic Text Document Repositories

[1]Wou Onn Choo, [3]Lam Hong Lee, [2]Dino Isa and [1]Wen Yeen Chue
[1]Taylor's Business School, Taylor's University, No. 1 Jalan Taylor's, 47500 Subang Jaya, Selangor, Malaysia
[2]Department of Electrical and Electronic Engineering, Faculty of Engineering,
University of Nottingham, Malaysia Campus, Jalan Broga, 43500, Semenyih, Selangor, Malaysia
[3]Intelligent Systems Research Group, Department of Electrical and Electronic Engineering,
Faculty of Engineering, University of Nottingham, Malaysia Campus,
Jalan Broga, 43500 Semenyih, Selangor, Malaysia

**Abstract:** This study proposes a folder allocation system for electronic text document thru the implementation of Bayesian classification approach. The manual folder allocation task for documents in computers has led to great inconveniences and extensive human efforts and involvements are required. This problem has become worse when the number of folders is huge and the structure is complex and continuously growing. In this study, an automatic folder allocation system is proposed by implementing Bayesian text classification approach. The Bayesian folder allocation system performs the tasks of managing, maintaining and organizing the documents in the storage space of a computer, by segmenting the documents into groups based on their topics, contents and similarities. From the experiments, Bayesian classifier has achieved the performance with the average accuracy of around 85%, with the highest up to 95.56% and lowest at 71.26%. The Bayesian folder allocation system has greatly minimized human involvement in allocating appropriate folder for documents, hence contributes to great convenience of the user in creating, editing, managing and storing electronic text documents in computer.

**Key words:** Bayesian approach, text classification, document folder alocation

## INTRODUCTION

There exists a growing phenomenon of textual data and the increasing availability of electronic documents due to the rapid growth of the usage of computers in creating, editing, transferring and storing documents in digital form. The task of automatically classifying documents into their requisite folders/directories is of great importance for managing, organizing and maintaining the information in the storage space of the computer. The storage space of a computer is needed to be managed, organized and maintained regularly. To carry out this task, extensive human efforts are required to determine and allocate the appropriate folder for the documents in the repositories. The same effort is required when new documents are wished to be stored in the computers. This situation brings great inconvenience to the computer users due to the iterative steps of folder allocation for documents. This will continue to worsen the problem when the number and layer of folders are expending over time and as the result, the structure of folders becomes complex with multiple layers. In this study, Bayesian classification approach is implemented to perform the task of automatic folder allocation by categorizing documents in the computer storage space, based on how the existing documents and folders were previously organized in the computers. This proposed automatic folder allocation system is able to recognize, differentiate and understand the contents of the existing documents and new incoming documents, hence determine their appropriate folders automatically. This could be done without the extensive commands and efforts from human. Nowadays, many of the processes in corporate, enterprises and even home users, have been automated to minimize the efforts and involvements of human. Human efforts have been well-known to be expensive, lack of consistency and permanence, less effective and less efficient. Therefore, the proposed Bayesian folder allocation system in this study contributes to a facility which greatly benefits the computer users in maintaining, managing and organizing text documents in computer storage space, in terms of efficiency and effectiveness.

The folder allocation system for electronic text document in computers is equipped with a text classifier to categorize documents into their annotated folders. Text

**Corresponding Author:** Lam Hong Lee, Intelligent Systems Research Group, Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Nottingham, Malaysia Campus, Jalan Broga, 43500, Semenyih, Selangor, Malaysia Fax: +603-89248017

document classification implies the declaration of a set of labelled categories to represent the documents, by using a statistical classifier trained with a labelled training set. In other words, the major task of a text classifier is to categorize text documents into one or more pre-defined categories. There are many ways to approach classification task. An increasing number of supervised classification approaches have been developed for documents such as decision tree induction (Greiner and Schaffer, 2001; Quinlan, 1993), rule induction (Apte *et al*., 1994a, b), k-nearest neighbor classification (Han *et al*., 1999), artificial neural network (Diligenti *et al*., 2003a, b), support vector machines (Joachims, 1998, 1999; Lin, 1999) and Bayesian classification (Domingos and Pazzani, 1997; Eyheramendy *et al*., 2003; Kim *et al*., 2002; McCallum and Nigam, 2003; Rish, 2001). Besides these, unsupervised clustering approaches such as k-means and self-organizing map, have also been utilized to segment text documents into clusters, based on their similarity, without requiring extensive prior knowledge.

Each of the approaches mentioned above has its own unique properties and associated problems. The decision tree induction algorithm and the rule induction algorithm are simple to understand and interpret. However, these algorithms do not work well when the number of distinguishing features between documents is large (Greiner and Schaffer, 2001; Quinlan, 1993). K-nearest neighbor algorithm is easy to implement and shows its effectiveness in a variety of problem domains (Han *et al*., 1999). As the major drawback, the k-NN algorithm is reported to be computationally intensive, and this problem is clearly seen with a large training set (Han *et al*., 1999). Artificial neural networks are outstanding with the ability in handling high dimensional documents with high level of noise and contradictory data. However, the high computational cost has become a major trade-off of this approach (Diligenti *et al*., 2003a, b). Support Vector Machines (SVM) can be used as a discriminative text document classifier and has been shown to be more accurate than most other techniques for classification tasks (Chakrabarti *et al*., 2003; Joachims, 1998). The main drawbacks of SVM text classification is the necessity in transforming text data into numerical format and its iterative binary classification processes in multi-class classification which consume a high computing time and cost.

Among the approaches mentioned above, Bayesian approach has become one of the most widely implemented machine learning classification techniques in various types of domains and applications. However it has been reported as a generative method which is relatively inaccurate than discriminative methods but less computational intensive in both of the training stage and the classifying stage (Brucher *et al*., 2002; Chakrabarti *et al*., 2003; Godbole, 2006; Joachims, 1998, 1999; Lin, 1999; Yang and Pedersen, 1997; Yang and Liu, 1999). On the other hand, Bayesian classification has been reported by some research works to be an approach which provides intuitive text generation models, hence it performs surprisingly well in many text domains, under some specific conditions (Domingos and Pazzani, 1997; Kim *et al*., 2002; McCallum and Nigam, 2003; Rish, 2001). Bayesian classifier requires only a small training set for the purpose of estimating the parameters' value, as it derives the precise nature of the probability model. Due to the assumption that Bayesian classification approach consists of independent variables it only requires the determination of the variances of variables for each of the categories but not the entire covariance matrix. Due to its nature of simple and low cost training and classifying algorithms, Bayesian classification technique has been widely implemented in variety of real world applications. Bayesian classification is widely utilized in applications which deal with textual data, such as spam e-mail filtering and e-mail categorization (Androutsopoulos *et al*., 2000; Guzella and Caminhas, 2009; Sahami *et al*., 1998; Xia *et al*., 2005).

By implementing Bayesian classification approach as the core framework of the proposed automatic folder allocation system for electronic text document in computers, the simple, low cost and high efficiency classification model is inherited from the nature of Bayesian classification. With the assistant of the proposed Bayesian classification framework, the folder allocation system is able to allocate the existing documents and incoming documents to their right folders automatically, without requiring extensive commands from human. This will greatly reduce human efforts and involvements in allocating the paths and folders to store the documents and improve the information retrieval and management processes.

## METHODOLODY

In this study, the tasks of managing, maintaining and organizing the documents in the storage space of a computer, by segmenting the documents into groups based on their topics, contents and similarities. The segmentation of documents based on their topics and similarities greatly contributes to the ease of searching existing documents, analyzing data and retrieving information and knowledge.

The classification task starts with the initial step of analyzing the text document and then "extracting" words which are contained in the document. To perform this analysis, a simple word extraction algorithm is used to extract each individual word from the document to generate a list of words. The list of words is constructed with the assumption that input document contains words w1, w2, w3,……, wn-1, wn, where the length of the document (in terms of number of words) is n.

The list of words is then used to generate a table, containing the probabilities of the word in each category which is related to the folder where the document can be stored. The column of "Word" is filled with words which are extracted from the input document. For the columns of probabilities of the particular word for each category, the values to be filled will be calculated by Bayesian classifier in the following stage. Table 1 illustrates the use of this method for the input document. Before Bayesian classifier performs the calculation of words' probability for each category it needs to be trained with a set of well-categorized training dataset. Each individual word from all training documents in the same category are extracted and listed in a list of words occurrence for that particular category, by using a simple data structure algorithm.

Based on the list of word occurrence, the trained Bayesian classifier calculates the posterior probability of the particular word of the document being annotated to a particular category by using the Bayesian equestion.

The equestion which is used to calculate the posterior probability of a word being annotated to a particular category is shown in Eq. 1:

$$\Pr(\text{category}\,|\,\text{Word}) = \frac{\Pr(\text{Word}\,|\,\text{Category}).\Pr(\text{Category})}{\Pr(\text{Word})} \quad (1)$$

The derived equation above shows that by observing the value of a particular word, wj, the prior probability Pr(Ci) of a particular category Ci, can be converted to the posterior probability, Pr(Ci|wj) which represents the probability of a particular word, wj being a particular category, Ci. The prior probability, Pr(Ci) can be compute from Eq. 2 and 3:

$$\Pr(Ci) = \frac{\text{Total\_of\_Words\_in\_Ci}}{\text{Total\_of\_Words\_in\_Training\_Dataset}} \quad (2)$$

$$= \frac{\text{Size\_of\_Ci}}{\text{Size\_of\_Training\_Dataset}} \quad (3)$$

Meanwhile, the evidence which call the normalizing constant of a particular word, wj, Pr(wj) is calculated by using Eq. 4:

$$\Pr(wj) = \frac{\sum \text{occurrence\_of\_wj\_in\_all\_categories}}{\sum \text{occurrence\_of\_all\_words\_in\_all\_categories}} \quad (4)$$

The total occurrence of a particular word in every category can be calculated by searching the training data base which is composed from the lists of word occurrences for every category. As previously mentioned, the list of word occurrence for a category is generated from the analysis of all training documents in that particular category during the initial training stage. The same method can be used to retrieve the sum of occurrence of all words in every category in the training data base.

To calculate the likelihood of a particular category, Ci with respect to a particular word, wj, the lists of word occurrence from the training data base is searched to retrieve the occurrence of wj in Ci and the sum of all words in Ci. These information will contribute to the value of Pr(wj |Ci ) given in Eq. 5:

$$\Pr(wi\,|\,Ci) = \frac{\text{Occurrence\_of\_wj\_in\_Ci}}{\sum \text{occurrence\_of\_all\_words\_in\_Ci}} \quad (5)$$

Based on the derived Bayes' Equestion for text classification, the posterior probability, Pr(Category|Word) of each word in the input document

Table 1: Table of words occurrences and probabilities used by the proposed bayesian classification approach to compute probability distribution of the input documents

| Word | Probability category 1 | Probability category 2 | Probability category 3 | --- | Probability category k-1 | Probability category k |
|---|---|---|---|---|---|---|
| w 1 | Pr (C1\|w1) | Pr (C2\|w1) | Pr (C3\|w1) | --- | Pr (C k-1\|w1) | Pr (C k\|w1) |
| w 2 | Pr (C1\|w2) | Pr (C2\|w2) | Pr (C3\|w2) | --- | Pr (C k-1\|w2) | Pr (C k\|w2) |
| w 3 | Pr (C1\|w3) | Pr (C2\|w3) | Pr (C3\|w3) | --- | Pr (C k-1\|w3) | Pr (C k\|w3) |
| - | - | - | - | - | - | - |
| - | - | - | - | - | - | - |
| - | - | - | - | - | - | - |
| w n-1 | Pr (C1\|wn-1) | Pr (C2\|wn-1) | Pr (C3\|w n-1) | --- | Pr (C k-1\|wn-1) | Pr (C k\|wn-1) |
| w n | Pr (C1\|wn) | Pr (C2\|wn) | Pr (C3\|wn) | --- | Pr (C k-1\|wn) | Pr (C k\|wn) |
| Total | ΣPr (C1\|W) | ΣPr (C2\|W) | ΣPr (C3\|W) | --- | ΣPr (C k-1\|W) | ΣPr (C k\|W) |
| Probability of input document | $\frac{\sum \Pr(C1\|W)}{n}$ | $\frac{\sum \Pr(C2\|W)}{n}$ | $\frac{\sum \Pr(C3\|W)}{n}$ | --- | $\frac{\sum \Pr(Ck-1\|W)}{n}$ | $\frac{\sum \Pr(Ck\|W)}{n}$ |

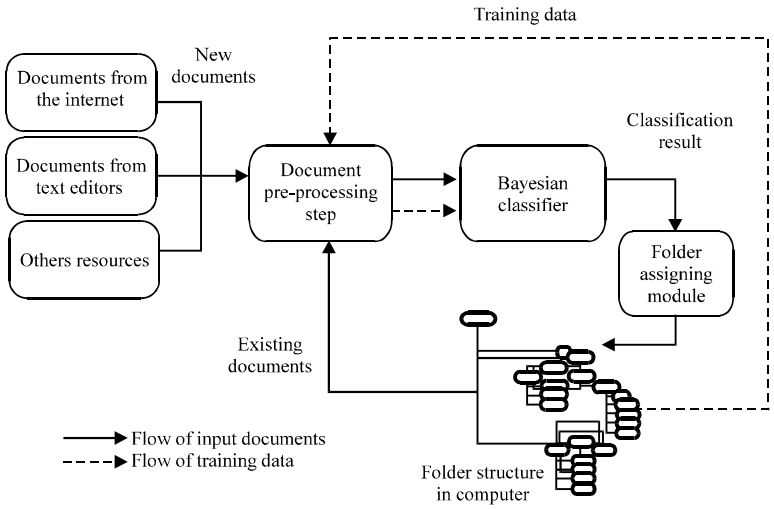


Fig. 1: Block Diagram of the Bayesian folder allocation system for document in computer

being annotated to each category can be measured. The calculation of Pr(Category|Word) is carried out by knowing the value of the prior probability Pr(Category), the likelihood Pr(Word|Category) and the evidence Pr(Word).

The posterior probability of each word being annotated to a category is then filled in at the appointed cells in the table as illustrated in Table 1. After all the "Probability" cells have been filled, the overall probability for an input document to be annotated to a particular category, Ci is calculated by dividing the sum of each of the "Probability" column with the length of the document, n which is shown in Eq. 6:

$$\Pr(Ci|Document)=\frac{\Pr(Ci\,|\,w1,\ w2,\ w3,.....,wn-1,wn)}{n} \quad (6)$$

where, w1, w2, w3,……, wn-1, wn, are the words which are extracted from the input document.

In the work presented here, Bayesian classifier is able to determine the right category/folder of an input document by referring to the probability values in the final row of Table 1, calculated by the trained classifier based on the Bayes equestion. The right category is represented by the one which has the highest posterior probability value, Pr(Category|Document), as stated in the Bayes classification rule.

## IMPLEMANTATION OF BAYESIAN CLASSIFICATION SCHEME FOR FOLDER ALLOCATION SYSTEM

It is the goal of this study is to implement the Bayesian classifier for the automatic folder allocation system for document in computers to allocate the documents in the folder structure based on their content. The documents to be classified could be the existing documents in the repositories, or new incoming documents which could be downloaded from the internet, or created using text editors, or from other resources such as the network grid.

In the context of automatic documents classification, a set of categories, C, is required. Each category represents either a subject or a discipline, C = {c1, c2, c3, ., cn}, where n is the number of categories in C. In the case of automatic folder allocation, C is the folders structure where documents are stored in the computer. In addition, D is defined as a set of new documents to be stored in the folder structure from variety of sources, D = {d1, d2, d3, ., dm}, where m is the number of documents in the collection. Automatic folder allocation is defined as a process in which a classifier determines to which category/folder a document belongs, based on its content. The main objective of classification is to assign an appropriate category/folder to a document with respect to the existing category set. Figure 1 illustrates the structure of the proposed Bayesian folder allocation system for documents in computer.

As for the training phase of the proposed automatic management system for document repositories, the folder structure in the document repositories which contains sample documents will be the training set for the management system for document repositories. The training process of the Bayesian classifier is carried out to train the management system to perform the task of categorizing documents into groups. The folder structure is initially designed and defined by the user and a small number of sample documents are categorized manually by the user to the right folder after an examination of their content. Since naïve Bayesian approach has been

implemented to perform the classification task, a large initial training set is not required. This feature is inherited from the nature and characteristic of the naïve Bayesian approach which assumes the individual attribute values as statistically independent. The naïve Bayesian classifier can be very robust to violations of its independent assumption. In other words, the naïve Bayesian classification approach is robust enough to ignore serious deficiencies in its underlying naïve probability model. The documents in the repositories may be represented in different formats such as Hyper Text Markup Language format (.html), Extensible Markup Language format (.xml), Adobe Portable Document Format (.pdf), Post Script format (.ps), Ms-Word document format (.doc and.docx) and Rich Text Format (.rtf). All the training documents will be transformed into a plain text file (.txt) format by a simple transformation process which is integrated in the proposed classification system. This is an advantage of the Bayesian classification, where the ability to handle raw text data is inherent to the process. In addition this is done without requiring any complicated transformation step to transform text data into representation suitable format, typically a numerical format, as required by most other classification approaches such as the support vector machines and the k-nearest neighbor. At the basic level, the Bayesian classifier examines a set of training documents that has been well-organized and categorized. Bayesian classifier then compares the content of all folders in the document repositories in order to build a list of words and their occurrences for each class/folder by using a simple data structure algorithm. These lists are used to identify or predict the membership of future documents to their right category, according to the probability of certain words occurring more frequently for certain categories. As the time goes on and more documents have been stored in the document repositories, the number of training documents increases and leads to the growth of the size of training data. Hence, the classifier is able to perform better as the training set grows. In the other words, the more the proposed document repositories management system works, the more intelligent it performs with higher accuracy (higher folder allocation accuracy).

After the training phase of the document repositories management system, the system is ready to perform its tasks of allocating documents into their right categories. The training set is acquired from a small amount existing documents in the repositories and manually organized by the human into groups of different topics based on the similarities and contents. There are still a large number of documents the repositories that are needed to be organized according to the topics of groups which have been labeled by the human. These documents are classified by the document repositories management system and stored according to their folders. As for the process of the automatic allocation of the new documents to the right folder, first, the system receives a new document which may be downloaded from the internet, or created using a text editor, or transferred via a grid network.

In the case that a document, di, that is wished to be store in the document repositories, a command is given by the user to the folder allocation system to examine the content of di. As similar to the training phase, di, needs to be transformed into the format of plain text file (.txt). After di has been transformed into the representation suitable format for the folder allocation system, a simple word extraction algorithm is used to extract each individual word from di prior to generating a list of words for di. Then, the Bayesian classifier will compute the posterior probability of di being annotated to each class/folder in the folder structure. The right class/folder of di can be determined by referring to the class/folder which has the highest posterior probability value, Pr(Class|di) among the other available class/folder in the folder structure, as stated in Bayes decision rule. In the other words, the system will allocate di to the most probable folder which contains similar documents in the document repositories.

## EXPERIMENTS AND EVALUATION

The prototype of the document repositories management system proposed in this study is developed by implementing the Bayesian classification technique as the core of the text documents allocation system. Therefore, the performance and accuracy of the Bayesian classifier can directly reflect the performance and accuracy of the proposed document repositories management system. In the experiments, a Bayesian classifier has been developed with Bayesian classification algorithm.

In order to simulate real world applications for the document repositories management system to be implemented in computers, five datasets have been acquired for experimental purpose to evaluate the performance of the proposed document repositories management system.

The Featured Articles dataset was designed and organized by the research group by extracting different types of articles from the Wikipedia website. A total of 1159 articles were acquired from 23 randomly selected categories. To build the training set, 10 documents from each category have been randomly selected. The remaining documents were prepared for testing purposes. In the other word, the training set contains 230 documents while the testing set contains a total of 929 documents. The list of selected categories for the Featured Articles dataset is shown in Table 2.

Table 2: List of categories of the featured articles dataset

| | |
|---|---|
| 1. Art, Architecture and Archaeology | 13. Literature |
| 2. Biology and medicine | 14. Mathematics |
| 3. Business, economics and finance | 15. Media |
| 4. Chemistry and mineralogy | 16. Music |
| 5. Computing | 17. Physics and astronomy |
| 6. Culture and society | 18. Politics and government |
| 7. Engineering and technology | 19. Religion and beliefs |
| 8. Geography and places | 20. Royalty, nobility and heraldry |
| 9. Geology, geophysics and meteorology | |
| 10. History | 21. Sport and games |
| 11. Language and linguistics | 22. Transport |
| 12. Law | 23. War |

Table 3: Grouped categories list of the 20-Newsgroups dataset

| | |
|---|---|
| comp.graphics | rec.autos |
| comp.os.ms-windows.misc | rec.motorcycles |
| comp.sys.ibm.pc.hardware | rec.sport.baseball |
| comp.sys.mac.hardware | rec.sport.hockey |
| comp.windows.x | |
| sci.crypt | misc.forsale |
| sci.electronics | |
| sci.med | |
| sci.space | |
| talk.politics.misc | talk.religion.misc |
| talk.politics.guns | alt.atheism |
| talk.politics.mideast | soc.religion.christian |

The Vehicles dataset was built by extracting vehicle related articles from the Wikipedia website. This dataset was acquired by extracting articles from 4 sub categories in the "Vehicles" category which are "Aircrafts", "Boats", "Cars" and "Trains". All 4 categories are easily differentiated and each category has a set of unique keywords. This dataset contains 640 documents. Each category contains 160 documents where 50 documents were used for training set and the remaining 110 documents were used for testing purposes. In other words, the training set contains a total of 200 documents with 50 documents for each category. The lists of categories of the Vehicles dataset are (1) Aircrafts, (2) Boats, (3) Cars and (4) Trains.

The Automobiles dataset is a dataset which was designed and organized by collecting articles about automobiles from the Wikipedia website. This dataset contains nine categories of automobiles, differentiated in terms of geographical regions and classification. There is a total of 30 documents have been acquired for each category with a sum total of 270 documents in the entire dataset. Out of the 30 documents, 10 documents from each category were extracted randomly to build the training set while the remaining 20 documents from each category make up the testing set. The lists of categories of the Automobiles dataset are (1) American Mini Vans, (2) American Sports Cars, (3) American SUVs, (4) Asian Mini Vans, (5) Asian Sports Cars, (6) Asian SUVs, (7) European Mini Vans, (8) European Sports Cars and (9) European SUVs.

A dataset containing articles about mathematical topics has been acquired from arxiv.org and is one of the datasets used to evaluation purpose. This dataset contains 8 mathematical sub-categories. There is a total of 40 documents for each category have been collected with a total of 320 documents in the entire dataset. There is a total of 10 documents from each category were extracted randomly to build the training set while the remaining 30 documents from each category were used for testing purposes. The lists of categories of the Mathematics dataset are (1) Algebraic Geometry, (2) Analysis of PDEs, (3) Combinatorics, (4) Differential Geometry, (5) Mathematical Physics, (6) Number Theory, (7) Probability and (8) Statistics.

The 20-Newsgroups dataset is one of the most common datasets used by many text classification research groups to evaluate the performance of their presented classification approaches. The 20-Newsgroups dataset is a collection of 20,000 Usenet articles from 20 different newsgroups with 1,000 articles per newsgroup. The 20-Newsgroups collection has become a widely-used dataset and is considered to be one of the standard benchmarks in datasets used for experiments in text applications of machine learning techniques, such as text classification and text clustering. The 20-Newsgroups dataset used in the experiments was acquired from the CMU Text Learning Group's website. In the experiments using this dataset, every category was divided into two subsets, in which 300 documents from each category were divided for training while the remaining 370 documents were used for testing purposes. In the other words, the training set contains 6,000 documents and the remaining 14,000 documents are used for testing purpose.

Table 3 illustrates the list of categories in the 20-Newsgroups dataset which have been categorized manually according to subject matter. By testing the proposed document repositories management system using the 20-Newsgroups dataset, a more generic evaluation of the accuracy of the proposed system can be conducted as compared to other ordinary or enhanced classification techniques.

Table 4 illustrates the comparison table for the classification accuracies of the proposed Bayesian document repositories management system, in handling different datasets. Experiments have been conducted on different datasets. However, the amount of training data and testing data for each experiment on different datasets are varying due to the availability and amount of documents in each dataset.

As illustrated in Table 4, the Bayesian classifier has achieved the performance with the average accuracy of around 85%. The experiments with most of the datasets

Table 4: Comparison On classification accuracies of bayesian classification on different datasets

| Data sets | Featured articles | Vehicles | Automobiles | Mathematics | 20 News groups |
|---|---|---|---|---|---|
| Size of training (Docs) | 230 | 200 | 90 | 80 | 6000 |
| Size of testing (Docs) | 929 | 440 | 270 | 240 | 14000 |
| Classification accuracy (%) | 71.26 | 95.56 | 87.22 | 88.75 | 82.57 |

have achieved the classification accuracy which is greater than 85%, where the experiment on Vehicles dataset has achieved 95.56% of accuracy which is the highest accuracy among the others. Meanwhile, the experiment on Mathematics dataset and the experiment on Automobiles dataset have achieved the classification accuracies of 88.75 and 87.22%, respectively. These classification accuracies have shown a promising performance of the implementation of the developed Bayesian classifier as the core classification framework for the proposed document repositories management system.

On the other hand, the experiment on 20-Newsgroups dataset has achieved the classification accuracy of 82.57% which is slightly lower than the average accuracy. This could be probably due to the number of categories of the 20-Newsgroups dataset is higher than the datasets which achieved good classification accuracy in their experiments, such as the Vehicles dataset, Automotives dataset and Mathematics dataset. With the nature of Bayesian classification, the classifier performs the computation by considering all the available categories together. This will lead to a greater confusion in differentiating all the categories in a single round, with a greater number of available categories in a classification task, hence degrade the performance of the classifier.

The experiment on the Featured Articles dataset achieved base line performance as compared to the rest of the experiments, with classification accuracy of 71.26%. This is probably due to the reason of the large number of available categories in the classification task, as similar to the experiment on the 20-Newsgroup dataset. Besides this, the ratio of training data and testing data may become one of the factors that affect the overall performance of the classifier. As the observation from the experiments on the ratio of training data and testing data of different datasets, most of the datasets have the training data to testing data ratio of more than 33-67. In other words, more than 33% of the entire dataset has been utilized for training purpose. However, the ratio of training data to testing data of the Featured Articles dataset is only 25-75. This ratio indicates the fact that only around 25% of documents from the Featured Articles dataset are utilized to train the classifier. Therefore, the experiment on the Featured Articles dataset has achieved relatively low classification accuracy as compared to the others.

## CONCLUSION

In conclusion, the presented automatic management system for electronic text document repositories has been proven to be effective and efficient as a tool to allocate the appropriate folders for documents. This system assists human in managing, maintaining and organizing text documents in computer for the ease of searching, analyzing and retrieving data, information and knowledge. This is achieved through the implementation of the simple, low cost but accurate Bayesian classification method to allocate the right category for the unlabeled documents in the repositories. The document repositories management system assists human in dealing with the huge and complex folder/directory structure when storing and organizing documents. It has successfully minimized human effort in allocating the appropriate folder for documents, thus contributes to the convenience of using computers to store, manage and analyze electronic documents. The experiments have shown that Bayesian classification contributed to good classification performance in most cases. However, for the cases with large number of available categories, such as the experiment on the Featured Articles dataset, low classification accuracy has been reported. This is due to the nature of Bayesian classification which considers all the categories in a single round while computing for the posterior probability values for a document to be annotated to each category. In overall, the performance of the proposed document repositories management system implemented with Bayesian classification technique is still promising with the great balance between the classification accuracy and the computational cost. This is also an important property of Bayesian classification scheme that causes it has been widely implemented in various domains of real world applications. For further improving the performance of the proposed document repositories management system, other potential classification schemes such as the support vector machines, has been investigated. The extension is planed to further reduce the computational cost and time, while maintaining the high classification accuracy of the folder allocation system for documents.

## REFERENCES

Androutsopoulos, I., J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. July 24-28, 2000, ACM, Athens, Greece, pp: 160-167.

Apte, C., F. Damerau and S.M. Weiss, 1994a. Automated learning of decision rules for text categorization. ACM Trans. Inform. Syst., 12: 233-251.

Apte, C., F. Damerau and S.M. Weiss, 1994b. Towards language independent automated learning of text categorization models. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 3-6, 1994, Dublin, Ireland, pp: 23-30.

Brucher, H., G. Knolmayer and M.A. Mittermayer, 2002. Document classification methods for organizing explicit knowledge. Technical Report, Research Group Information Engineering, Institute of Information Systems, University of Bern, Switzerland.

Chakrabarti, S., S. Roy and M.V. Soundalgekar, 2003. Fast and accurate text classification via multiple linear discriminant projections. VLDB J., 12: 170-185.

Diligenti, M., M. Maggini and L. Rigutini, 2003a. Automatic text categorization using neural networks. Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, November 2, 1996, Washington, DC., USA., pp: 59-72.

Diligenti, M., M. Maggini and L. Rigutini, 2003b. Learning similarities for text documents using neural networks. Proceedings of the Artificial Neural Networks in Pattern Recognition (ANNPR), September 12-13, 2003, University of Florence italy, pp: 1-6.

Domingos, P. and M. Pazzani, 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn., 29: 103-130.

Eyheramendy, S., A. Genkin, W.H. Ju, D. Lewis and D. Madigan, 2003. Sparce bayesian classifiers for text categorization. Technical Report, Department of Statistics, Rutgers University. http://www.stat.r utgers.edu/~madigan/ PAPERS/ jicrd-v13.pdf

Godbole, S., 2006. Inter-class relationship in text classification. Ph.D. Thesis, Indian Institute of Technology, Bombay, India.

Greiner, R. and J. Schaffer, 2001. AIxploratorium: Decision trees. Department of Computing Science, University of Alberta, Edmonton, AB T6G 2H1, Canada. http://www.cs.ualberta.ca/~aixplore/ learning/DecisionTrees

Guzella, T.S. and W.M. Caminhas, 2009. A review of machine learning approaches to spam filtering. Expert Syst. Appl., 36: 10206-10222.

Han, E.H., G. Karypis and V. Kumar, 1999. Text categorization using weight adjusted k-nearest neighbour classification. Technical Report, Department of Computer Science and Engineering, Army HPC Research Center, University of Minnesota. http://glaros.dtc. umn.edu/gkhome/fetch/ papers/wknnPAKDD01.pdf

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Springer, Berlin, Heidelberg, pp: 137-142.

Joachims, T., 1999. Making Large-Scale SVM Learning Practical. In: Advances in Kernel Methods-Support Vector Learning, Scholkopf, B., C.J.C. Burges and A.J. Smola (Eds.)., MIT Press, Cambridge, MA., ISBN-10: 0-262-19416-3, pp: 169-184.

Kim, S.B., H.C. Rim, D.S. Yook and H.S. Lim, 2002. Effective methods for improving naive bayes text classifiers. Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence, Volume 2417, August 18-22, 2002, Tokyo, Japan, pp: 414-423.

Lin, Y., 1999. Support vector machines and the bayes rule in classification. Technical Report No. 1014, Department of Statistics, University of Wiscousin, Madison. http://www.stat.wisc. edu/public/ftp/yilin /tr1014.ps

McCallum, A. and K. Nigam, 2003. A comparison of event models for naive bayes text classication. J. Mach. Learn. Res., 3: 1265-1287.

Quinlan, J.R., 1993. C4.5: Program for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA., USA., ISBN: 1-55860-238-0, Pages: 302.

Rish, I., 2001. An empirical study of the naive bayes classifier. Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, August 4, 2001, Seattle, USA., pp: 41-46.

Sahami, M., S. Dumais, D. Heckerman and E. Horvitz, 1998. A bayesian approach to filtering junk E-Mail. Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, July 26-27, 1998, Madison, Winconsin, pp: 55-62.

Xia, Y., W. Liu and L. Guthrie, 2005. Email categorization with tournament methods. Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, June 15-17, 2005, Alicante, Spain, pp: 150-160.

Yang, Y.M. and J. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning, July 8-12, 1997, Nashville, TN., USA., pp: 412-420.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization method. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, New York, USA., pp: 42-49.