TEXT CATEGORIZATION USING NAIVE BAYES ALGORITHM

Wan Hazimah binti Wan Ismail @ W. Abdullah
Mrs. Siti Sakira binti Kamaruddin
Mr. Mohd Shamrie bin Sainin

Faculty of Information Technology
University Utara Malaysia
06010 Sintok, Kedah
Malaysia

Email: whazi81@yahoo.com, sakira@uum.edu.my, shamrie@uum.edu.my

Abstract

As the volume of information available on the internet and corporate intranet continues to increase, there is a growing interest in helping people better find, filter, and manage all these resources. Text categorization is one of the techniques that can be applied in this situation. This paper presents text categorization system based on naive Bayes algorithm. This algorithm has long been used for text categorization tasks. Naive Bayes classifier is based on probability model that integrate strong independence assumptions which often have no bearing in reality. The aims of this project are to categorize the textual document using naïve Bayes algorithm and to measure the correctness of the chosen technique for the categorization process. This paper also discusses the experiment in categorizing articles using naive Bayes.

1. INTRODUCTION

Information breaks into two broad categories which is structured and unstructured. Structured data is the data that can obtain in databases which every bit of information has an assigned format and significance. Unstructured data is what we find in emails, reports, PowerPoint presentations, voice mail, phone notes, agendas and photographs.

Managing unstructured data is a problem that has been around since people using computers to write letters, reports, and send emails. Moreover, the huge amount of unstructured data that are available on the web and intranets also creates an information overloading problem. According to Robb (2004) "We are drowning in information but are starving for knowledge". The information is only useful when it can be located and then synthesized into knowledge. Because of this problem, managing the unstructured data that contained in textual documents is important. The unstructured data usually have a free format, mostly in text form, which are difficult to manage. Therefore, many researchers have proposed various techniques to arrange such unstructured data effectively.

Organizers of conference which deal with unstructured data also faced problem that related to document management. The main problem is to organize the text documents into manageable and easy to understand categories. Text categorization today is a necessity due to the very large amount of text documents that have to deal with daily. This technique has become one of the key techniques for handling and organizing text data. According to Basu, Watters, and Shepherd (2002) text categorization is the process of assigning documents into a fix number of predefined categories or classes based on their content. In 2003, Rao state that "a categorization system creates and maintains a hierarchical structure of categories". Text categorization system can be used in indexing documents to assist information retrieval tasks as well as classifying documents. Based on Fei et al. (2004) they state that "text categorization technology will gradually combine with some information processing technologies such as search engine and information filtration, which will improve the quality of information service effectively".

Most of the existing approaches for text categorization are based on statistics and machine learning. According to Mitchell (1997), machine learning is computer programs that learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if it's performance at tasks in T, as measured by P, and improves with experience E. In general, machine learning is about learning to do better in the future base on what was experienced in the past. Naïve Bayes is one of the techniques in machine learning and it is used to categorize text documents.

Naive Bayes

Naive Bayes is based on Bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. It often called as naïve Bayes classifier. Based on Wikipedia (2005), "Bayesian is the philosophical tenet that the mathematical theory of probability applies to the degree of plausibility of a statement". Bayesian probability was proposed by Thomas Bayes who proved a special case which is Bayes's theorem. Bayesian probability comes into use in 1950. Naive Bayes classifier is a simple probability classifier. It computes the likelihood that a program is malicious given the features that are contained in the program ("Naïve Bayes", 2001). Naive Bayes classifier also is a simple algorithm for categorization process and gives a good performance in getting a result (Shen & Jiang, 2003).

The basic text categorization in forum message application has been discussed in Sainin (2005). The paper explains about the study of the naive Bayes algorithm to classify forum messages whether clean or bad, where clean messages has no bad words, while bad messages contains one or more bad words. The performance in the forum application can be considered high. Therefore, this study was using the algorithm to classify conference paper.

1.1 Problem Statement

With the recent developments in technologies now, managing the information is become very important especially information in unstructured form. Such unstructured data are text documents, emails, images, audio, and video data. It is difficult to retrieve the useful data from the unstructured data.

Organizers of conference today, faced the key issues in organizing the submitted proceeding articles. The proceeding articles are based on various field of Information Technology such as Knowledge Management, Artificial Intelligence, Networking, and others. All these documents are classified into categories manually by scanning one's contents. Therefore it consumes time to classify documents manually. This factor was also effect the process of managing the conference session which is time consuming. Other than that, the searching and browsing process for the needed articles take more time. This is because the articles are not well organized. Sometimes because of this problem, some articles maybe lost or misplace.

1.2 Project's Objectives

The objectives of this study are specified as follows:

i. To categorize the textual document using naïve Bayes technique.
ii. To measure the correctness of the chosen technique for the categorization process.

1.3 Project's Scope

The scope of this study is as below:

i. The unstructured data to be categorized is a sample of proceeding articles that are in text format. The sample of the proceeding articles were obtained from the Knowledge Management International Conference (KMICE) 2004. These proceeding articles consist of various fields in Information Technology.
ii. In this study, only the abstract of the proceeding articles were used. There are six (6) categories of this proceeding articles which are Artificial Intelligence, Concept, E-learning, Engineering, Survey and Case Study, and System.
iii. The total article in the KMICE 2004 is 87 articles. Based on this study, 80% from the articles was used for training process while another 20% for testing process.
iv. Besides that, the scope of this study is to categorize the proceeding articles by using naive Bayes technique.

2. LITERATURE REVIEW

Unstructured data is an unstructuredness data that are in free format and usually in a text form. Such unstructured data are emails, html texts, images, video, and text documents (Adam & Gangopadhyay, 1998; Chaovalit & Zhou, 2005; Ishikawa et al., 1998). Usually this unstructured data is difficult to manage and time-consuming to retrieve information because the data is not well defined. Unstructured data is a large volume of data compared to structured data and it is also known as bulk data or BLOB (binary large objects) (Ishikawa et al., 1998). Structured data is information that has been organized to allow identification and separation of the context of the information from its content. It is uniform, consistent pattern for arranging data in a file. When data is structured in a file, the data can easily used with less chance for errors. The unstructured data is more valuable data compared to structure data for the organization. This is because more information can be obtained from the documents, emails, images, and others which is the type of unstructured data.

Text categorization is the process of assigning documents into a fix number of predefined categories or classes based on their content (Chen, Zhou, & Wu, 2004; Dong & Han, 2004; Dumais et al., n.d; Joachims, 1998; Lin & Hu, 2004; Peng, Schuurmans, & Wang, 2003). This is the goal of text categorization. Another goal is text in a text categorization may be relevant to one or more categories corresponding to single-label or multi-label categorization problem (Chen et al., 2004).

In this project, machine learning approach is applied to text categorization. This approach is suitable for text categorization. Based on Bengio (2005), learning means changing in order to be better when a similar situation arrives. It is also an essential of human property. According to Mitchell (1997), machine learning is a multidisciplinary field that brings together scientist from artificial intelligence, computational complexity theory, probability and statistics, information theory, philosophy, and neurobiology. Machine learning also is a core subarea of artificial intelligence.

According to Schapire (2003), machine learning learns computer algorithms for learning to do some task. This type of learning is a learning to complete a task, or to make accurate predictions, or to behave intelligently. The learning is based on some sort of observations or data, such as examples, direct experience, or instruction. Generally, machine learning is about learning to do better in the future base on what was experienced in the past.

Naive Bayes is the technique that been chosen for categorizing text. It is a technique that based on probability models that incorporate strong independence assumptions which often have no bearing in reality, hence are naive. Bayesian is the mathematical theory of probability that applies to the degree of plausibility of a statement. This statement is based on belief of philosophical. The Bayesian interpretation of probability allows probabilities to be assigned to random events and it also allows the assignment of probabilities to any other kind of statement. The Bayesian probability or Bayesian theory is presented by Thomas Bayes who has been proved a special case which is Bayes' theorem ("Bayesian probability", 2005). The naive Bayes algorithm is depicted in Figure 2.1

LEARN_NAIVE_BAYES (*Examples*)

Examples is a set of text documents along with their target values which V is the set of all possible target values. This function learns the probability terms $P(w_k | v_j)$, and can also learn the class prior probabilities $P(v_j)$.

```
• Calculate the required $P(v_j)$ and $P(w_k | v_j)$ probability terms
      For each target value $v_j$ in V do
            - $docs_j \leftarrow$ the subset of documents from Examples for which the
                  target value is $v_j$
            - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
            - $Text_j \leftarrow$ a single document created by concatenating all
            members of $docs_j$
        - $n$ ← total number of distinct word positions in $Text_j$
        - For each word $w_k$ in Vocabulary
            $n_k$ ← number of times word $w_k$ occurs in $Text_j$
            $P(w_k | v_j) \leftarrow \frac{n_k + 1}{n + | Vocabulary |}$

CLASSIFY_NAIVE_BAYES_TEXT (Text)
Returns the estimated target value for the document Text. $a_i$ is the word found in the ith
position within Text.

    • positions ← all word positions in Text that contain tokens found in Vocabulary
    • Return $V_{NB}$ where
            $V_{NB} = \arg\max_{vj \in V} \prod_i P(a_i | v_j)$
```

Figure 2.1: Naive Bayes Algorithm

Naive Bayes is a simple probabilistic classifier and have been widely used for the probabilistic text categorization. Moreover, it achieves comparable performance with popular classifiers and constantly outperforms competing algorithms which is based on the experiments that has been done by the past researchers (Huang & Hsu, 2002; Kim, Rim, & Lim, 2002). The naive part for this method is the assumption of word independence for example the conditional of a word given a category is assumed to be independent from the conditional probabilities of other words given that category (Yang & Liu, 1999). It can be trained efficiently in a supervised learning setting ("Naive Bayes classifier", 2005). The target of the Bayesian classification is to decide and choose the class that maximizes the posteriori probability (Wang et al., 2004).

Based on research that been done by Phung, Bouzerdoum, Chai, and Watson (2004), naive Bayes can also been applied in face detection application. Naive Bayes face/nonface classifier is been used in the face detection technique. This classifier is been used to analyze the effect on classification performance of preprocessing, feature extraction schemes, and classifier combination techniques.

Provost (1999) has done a research in comparing two learning algorithms which is naive Bayes that used for text categorization and RIPPER which is a rule-learning approach that used for email categorization. RIPPER is highly suitable for email classification and filtering whereas naive Bayes advantage in classification accuracy. Based on the three experiments that been done which are hand-sorted mail, automatically-sorted mail, and spam detection shows that naive Bayes outperforms RIPPER in classification accuracy

According to Wu et.al (2002), the performance of Rocchio classifier and naive Bayes classifier can be improved by using refinement model. This approach is able to improve text categorization accuracy. These two techniques are also suitable for very large text collections. It is because these two techniques allow the data to reside on disk and need only one scan of the data to build a text classifier.

Boosting algorithm is the algorithm that "builds a mapping from category document pairs to real-valued scores and uses the scores to provide a ranking of categories" (Chen et al., 2004). Boosting algorithm is been used to find a highly accurate categorization rule by combining many weak hypotheses and it is also maintains a set of importance weights over training examples and labels. Chen et.al (2004) has been adopted this technique for categorizing a multi-label Chinese documents.

Another technique is maximum entropy which is an estimation technique for probability distribution from data. Besides text categorization, this technique can also be used in language modeling, part-of-speech tagging, and text segmentation. Nigam, Lafferly, and McCallum (1999) used this technique for text categorization by estimating the conditional distribution of the class variable given the document. In this experiment, it shows that sometimes maximum entropy is significantly better and sometimes become worse.

3. RESEARCH METHODOLOGY

According to Bennet et al. (2002), a methodology is a set of principles that guide a practitioner and manager to choose a particular method suited to a specific project. The research methodology is adopted from the Design Research in Information Systems (IS) which is provide with useful information that based on understanding, evaluating, and publishing design research. This research methodology is based on Information Systems (IS) that been proposed by Hevner et al. (2004). There are five phases in this research methodology which are awareness of problem, suggestion, development, evaluation, and conclusion. The phases can be depicted in Figure 3.1 below.

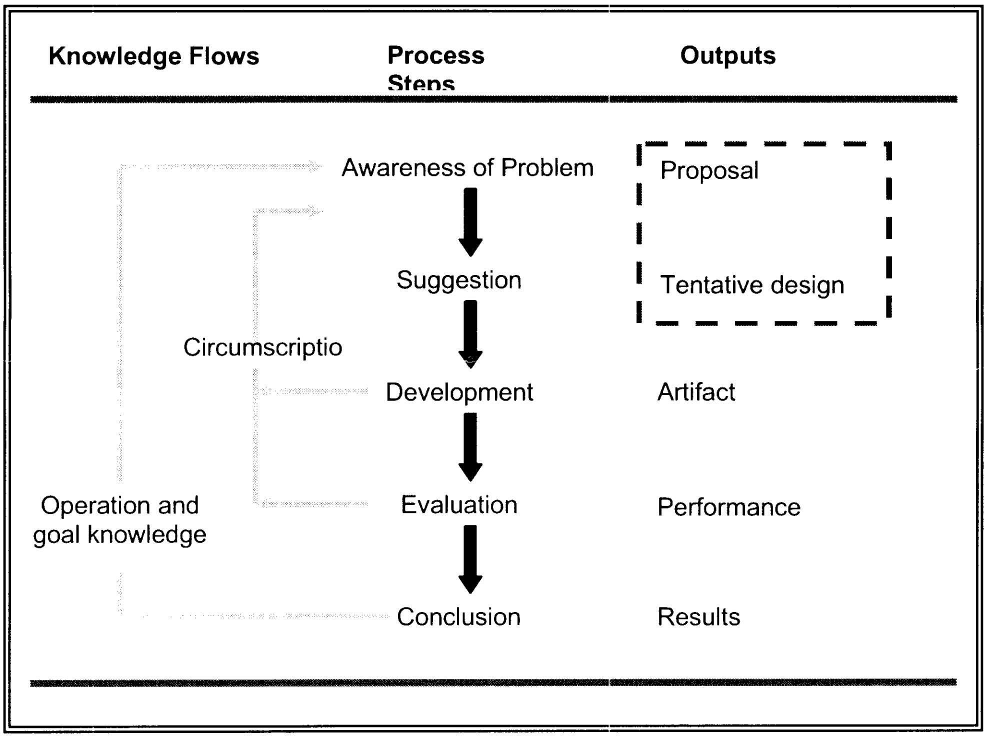

Figure 3.1: The General Methodology of Design Research

Phase 1: Awareness of Problem

The first phase in this methodology is the awareness of problem. This phase is a phase that describe about an awareness of an interesting problem that may come from multiple sources. In this study the scope of the problem is based from the conference committee problem. The key issue that the conference committee faced is to organize the submitted proceeding articles into categories automatically. The output of this phase is a proposal, formal or informal, for a new research effort.

Phase 2: Suggestion

After identifying the problems in phase one, all the problems are analyzed to find a best solution. All the possible solution for the problems are listed out and the pros and cons of each solution are identified. Based on the results, conference committee members find that the organization should develop a system that can manage all the textual documents in a meaningful way. They tend to develop a prototype system first, in order to measure the accuracy of the system.

The first task is to get a set of data. The data is the proceeding articles that were obtained from the Knowledge Management International Conference (KMICE) 2004 collections. The total articles in this collection of proceeding articles are 87. This data is divided for two process which are training (66 articles), and testing (21 articles). After all the data has been collected, the next task is to preprocess all the data. In this process, the articles abstract were extracted in order to get distinct words. All the characters in the selected articles are categorized independently which means the punctuation marks, digits, parenthesis, etc is been removed. Then, model for classification was built based on the probability. In order to build this model, the naive Bayes algorithm was applied. The final process is to design the prototype system for text categorization. Figure 3.2 shows the architecture of the prototype system. The design is based on the Unified Modeling Language (UML). The tool that was used to design the prototype system is Rational Rose 2000.

Phase 3: Development

After producing the solid tentative design from the suggestion phase, the system was developed. The development of the prototype system is based on the specified requirement that was identified in previous phase. The development of the prototype is based on naïve Bayes algorithm. This algorithm is one of the text categorization algorithms and it can perform efficiently in classifying text. In this project, the prototype system was developed using Visual Basic (VB) programming language.

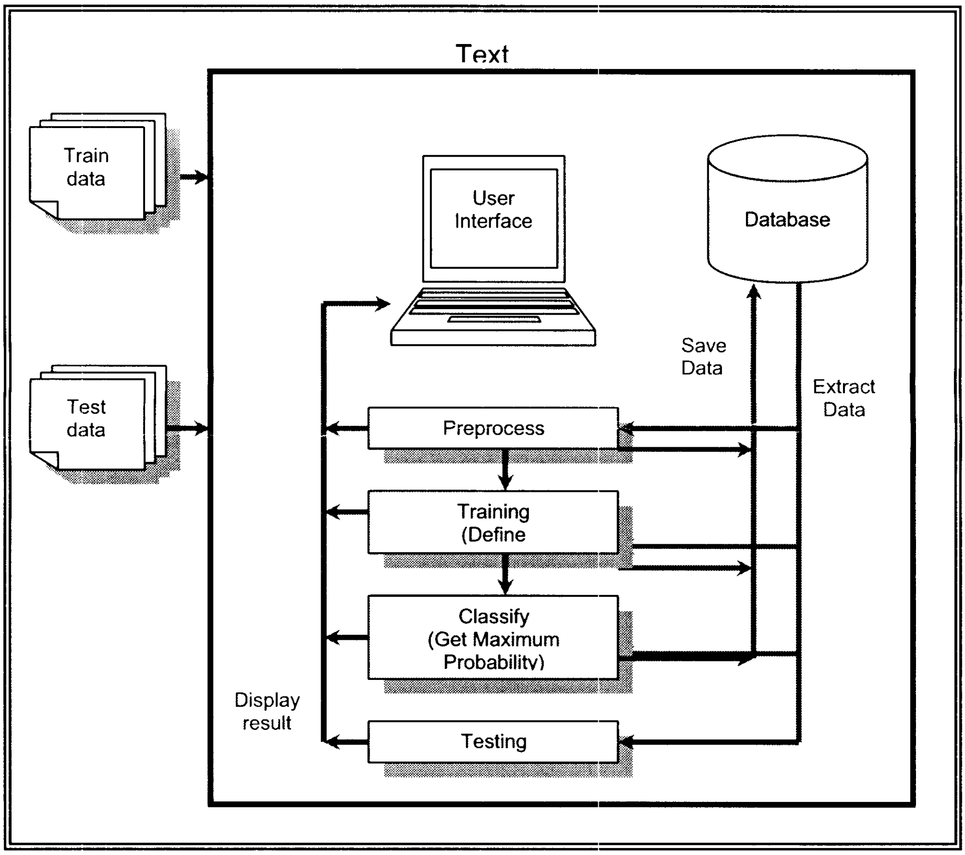

Figure 3.2: The Architecture for Text Categorization Prototype System

Phase 4: Evaluation

In this phase, the evaluation process is focuses on two elements which is naive Bayes algorithm and the prototype system that was developed in the previous phase (development). This phase also is the phase that measures the performance of the prototype system. It is based on the accuracy in categorizing the articles. The accuracy was evaluated using confusion matrix. In evaluating the prototype system, the results from the conducted research are compared with the categorization done manually by the experts. The maximum probability was identified in order to assign the category for each article.

Phase 5: Conclusion

This phase is the finale of a specific research effort. The deliverable for this phase is the result that was obtained from the text categorization prototype system.

4. RESULTS

The result was divided into two categories which are the result that based on training articles (66 articles) and the unseen articles (21 articles) that was used for testing in order to validate the performance of the algorithm. Based on this experiment, there are a total of 2458 words were produced and this word was saved in a database. The probability value for each word in each category was obtained in the learning process. The classification process is to assign the category of each article automatically. The maximum probability that was obtained from the experiment can be assigned as a result for categorization.

Confusion matrix was used to depict the output from the prototype system. The performance of the naive Bayes algorithm was evaluated using the data in the matrix. Based on the calculation, the result shows that naive Bayes produce 81.82 percent (%) accuracy for training articles. This can be shows in the confusion matrix in Table 4.1.

Table 4.1: Confusion Matrix for Training Articles

Prediction Actual	Artificial Intelligence	Concept	E-learning	Engineering	Survey & Case Study	System	Total
Artificial Intelligence	8	1	0	2	1	0	12
Concept	0	15	0	0	1	0	16
E-learning	1	1	10	0	0	0	12
Engineering	0	0	0	0	2	0	2
Survey & Case Study	1	1	0	0	16	0	18
System	0	0	0	0	1	5	6

The result shows that none of the category classified the entire article correctly. There are 12 articles out of 54 articles that are incorrectly classified. For Artificial Intelligence (AI) category, eight articles were correctly classified while another four articles are incorrectly classified. Only one of the articles in concept category was not classified correctly. The article was classified as survey and case study category. This situation is same for system category where one of the articles was classified as survey and case study. For E-learning category, there are 10 articles were classified correctly. There are none of the articles that were correctly classified in engineering category. The articles that were correctly classified in survey and case study category are 16 articles out of 18 articles. Overall, the articles that are correctly classified are 54 (8+15+7+10+16+5) and the training accuracy was computed as follows.

Training Accuracy = 54/66 * 100
= **81.82 %**

The summarization for classification process based on training articles is shown in Table 4.2 as follows.

Table 4.2: Summarization of Classification (Training)

Category	Correct	Incorrect
Artificial Intelligence	8	4
Concept	15	1
E-learning	10	2
Engineering	0	2
Survey & Case Study	16	2
System	5	1

To validate the performance of the naïve Bayes algorithm, it was then applied to the testing articles. There are 21 articles in the testing process which also based on KMICE 2004. Based on the calculation, the accuracy achieved using the algorithm was 47.62 percent (%). The result is depicted in Table 4.3.

Table 4.3: Confusion Matrix for Testing Articles

Prediction **Actual**	**Artificial Intelligence**	**Concept**	**E-learning**	**Engineering**	**Survey & Case Study**	**System**	**Total**
Artificial Intelligence	3	0	0	0	0	0	3
Concept	0	2	0	0	3	0	5
E-learning	0	0	2	0	2	0	4
Engineering	0	1	0	0	0	0	1
Survey & Case Study	2	0	0	0	3	0	5
System	0	1	0	0	2	0	3

Based on table 4.2 above, it shows that only some of the articles were correctly classified. All the articles from the AI category were classified correctly. Two articles from the concept category were classified correctly whereas the other articles were incorrectly classified as survey and case study. For E-learning category, two articles were correctly classified while another two articles were classified as survey and case study category. Engineering category was incorrectly classified the article where the article was classified as concept category. Two articles out of five articles were classified as AI in survey and case study category. There are none of the articles that were correctly classified in the system category. Overall, the articles that correctly classified are 10 articles (3+2+2+3) out of 21 articles. The testing accuracy was computed below.

Testing Accuracy = 10/21 * 100
= **47.62 %**

The summarization for classification process based on testing articles is shown in Table 4.4 below.

Table 4.4: Summarization of Classification (Testing)

Category	Correct	Incorrect
Artificial Intelligence	3	0
Concept	2	3
E-learning	2	2
Engineering	0	1
Survey & Case Study	3	2
System	0	3

Based on this result, the accuracy that was achieved is lower. It is because there is not enough vocabulary to classify the articles correctly. Naive Bayes classification is dependent to the probability of attributes. If more words are ignored, then the classification accuracy towards certain category is low. This problem can be overcome by training with more words in order to provide larger space for classification.

In this classification process, the minimum probability boundary was applied in order to avoid the probability value zero in this prototype system (Visual Basic). The minimum floating point that been used in this project is 4.94065645841247E-320.

5. SIGNIFICANCE

The result from this study provides solution for organizing committee in categorizing the submitted documents automatically. By automatically categorizing the documents, it is easier to organize the session for the conference. This factor will reduce cost and time for the organization. Besides that, the process for searching and browsing the documents become more effective and efficient. This study also gives an impact to the management of the documents which is better than the manual categorization process.

6. CONCLUSION

Naive Bayes classifier is widely used in many classification tasks because its performance is competitive. It is also simple to implement and it possesses fast execution speed. This study presented an approach in categorizing the proceeding articles using the naïve Bayes algorithm. Based on the experiment, naive Bayes is suitable in categorizing a large set of data. It must have enough vocabulary in order to obtain a higher classification. The classification accuracy for training articles outperform better than testing articles.

This experiment is based on the abstract of the proceeding articles, in the future enhancement, the prototype system will categorize based on the whole content of articles. Besides that, the functionality and the performance of the system will be enhanced in order to make the system applicable to the conference application.

REFERENCES

Adam, N., R., & Gangopadhyay, A. (1998). Content-based retrieval in digital libraries. *Technical Activities Forum.* Retrived June 20, 2005, from IEEE Xplorer database.

Bayesian probability. (2005, June 22). Wikipedia Encyclopedia. Retrieved August 3, 2005 from http://en.wikipedia.org/wiki/Bayesian_probability.

Bengio, S. (2005). Statistical machine learning. Retrieved August 14, 2005 from http://www.idiap.ch/~bengio/lectures/intro.pdf.

Chaovalit, P., & Zhou, L. (2005). Movie review mining: a comparison between supervised and unsupervised classification approaches. Proceedings *of the 38th Hawaii International Conference on System Sciences.* Retrived June 20, 2005, from IEEE Xplorer database.

Chen, J., Zhou, X., & Wu, Z. (2004). A multi-label Chinese text categorization system based on boosting algorithm. *Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04).* Retrieved August 2, 2005 from IEEE Xplorer database.

Dong, Y., S., & Han, K., S. (2004). A comparison of several ensemble methods for text categorization. *Proceedings of the 2004 IEEE International Conference on Services Computing (SCC'04).* Retrieved August 6, 2005 from IEEE Xplorer database.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (n.d). Inductive learning algorithms and representations for text categorization. Retrieved July 23, 2005 from http://research.microsoft.com/~sdumais/cikm98.pdf.

Fei, Y., Jiyao, A., Hong, L., Miaoling, Z., & Ouyang, Y. (2004). Intelligence text categorization based on Bayes algorithm. *Proceedings of 2004 International Conference on Information Acquisition.* Retrieved July 6, 2005, from IEEE Xplorer database.

Hevner, A., March, S., Park, J. & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly *28*(1).

Huang, H., J., & Hsu, C., N. (2002). Bayesian classification for data from the same unknown class. *IEEE Transaction on System, Man, and Cybernetics.* Retrieved July 18, 2005 from IEEE Xplorer database.

Ishikawa, H., Kubota, K., Noguchi, Y., Kato, K., Ono, M., Yoshizawa, N., & Kanaya, A. (1998). A document warehouse: a multimedia database approach. Retrived June 20, 2005, from IEEE Xplorer database.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. Retrieved July 27, 2005 from http://www.cs.cornell.edu/People/tj/publications/joachims_98a.pdf

Kim, S., B., Rim, H.,C., & Lim, H., S. (2002). A new method of parameter estimation for multinomial naïve Bayes text classifiers. Retrieved August 16, 2005 from ACM Digital Library database.

Lin, J., H., & Hu, T., F. (2004). Fuzzy correlation and support vector learning approach to multi-categorization of documents. *2004 IEEE International Conference on Systems, Man, and Cybernetics.* Retrieved July 28, 2005 from IEEE Xplorer database.

Mitchell, T., M. (1997). *Machine Learning.* McGraw Hill, New York: NY.

Naive Bayes classifier. (2005, July 5). Wikipedia Encyclopedia. Retrieved August 3, 2005 from http://en.wikipedia.org/wiki/Naive_Bayes.

Nigam, K., Lafferty, J., McCallum, A. (1999). Using maximum entropy for text classification. Retrieved August 23, 2005 from http://www.cs.cmu.edu/People/knigam/papers/maxent-ijcaiws99.pdf.

Peng, F., Schuurmans, D., & Wang, S. (2003). Language and task independent text categorization with simple language models. *Proceedings of HLT-NAACL*. Retrieved August 2, 2005 from IEEE Xplorer database.

Phung, S., L., Bouzerdoum, A., Chai, D., & Watson, A. (2004). Naïve Bayes face/nonface classifier: a study of preprocessing and feature extraction techniques. *2004 International Conference on Image Processing (ICIP)*. Retrieved August 6, 2005 from IEEE Xplorer database.

Provost, J. (1999). Naive Bayes vs. rule-learning in classification of email. Retrieved August 16, 2005 from http://www.cs.utexas.edu/users/jp/research/publications/provost-ai-tr-99-281.pdf.

Robb, D. (2004, Sept 13). Getting the bigger picture: dealing with unstructured data. Retrieved August 20, 2005 from http://www.enterpriseitplanet.com/storage/features/article.php/3407161.

Sainin, M. S. (2005). Applying Learning to Filter Text in Forum Message. In Proceeding, Socio-Economy & Information Technology Seminar 3, UUM, Perlis.

Schapire, R. (2003). Foundations of machine learning. Retrieved August 14, 2005 from http://www.cs.princeton.edu/courses/archive/spring03/cs511/scribe_notes/0204.pd f

Shen, Y., & Jiang, J. (2003). Improving the performance of naive Bayes for text classification. Retrieved

August 1, 2005, from http://nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf

.

Wang, L., M., Yuan, S., M., Li, L., & Li, H., J. (2004). Boosting naive Bayes by active learning. *Proceedings of the Third International* Conference *on Machine Learning and Cybernetics*. Retrieved July 29, 2005 from IEEE Xplorer database.

Wu, H., Phang, T., H., Liu, B., & Li, X. (2002). A refinement approach to handling model misfit in text categorization. Retrieved July 28, 2005 from ACM Digital Library database.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. Retrieved August 1, 2005, from ACM Digital Library database.

Zadok, E. (2001). Naive Bayes. Retrieved August 1, 2005 from http://www.fsl.cs.sunysb.edu/docs/binaryeval/node5.html.