

**TEXT CATEGORIZATION USING NAIVE BAYES  
ALGORITHM**

**WAN HAZIMAH BINTI WAN ISMAIL**

**UNIVERSITI UTARA MALAYSIA (2005)**

# **TEXT CATEGORIZATION USING NAIVE BAYES ALGORITHM**

A thesis submitted to the Faculty of Information Technology in  
partial fulfillment of the requirements for the degree Master of Science (Information  
Technology), Universiti Utara Malaysia

By

Wan Hazimah binti Wan Ismail

Copyright© Wan Hazimah binti Wan Ismail, 2005. All rights reserved.



**JABATAN HAL EHWAL AKADEMIK**  
*(Department of Academic Affairs)*  
**Universiti Utara Malaysia**

**PERAKUAN KERJA KERTAS PROJEK**  
*(Certificate of Project Paper)*

Saya, yang bertandatangan, memperakukan bahawa  
*(I, the undersigned, certify that)*

**WAN HAZIMAH BINTI WAN ISMAIL**

calon untuk Ijazah  
*(candidate for the degree of)* **MSc. (IT)**


telah mengemukakan kertas projek yang bertajuk  
*(has presented his/her project paper of the following title)*

**TEXT CATEGORIZATION USING NAIVE BAYES ALGORITHM**

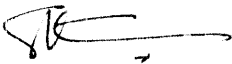
seperti yang tercatat di muka surat tajuk dan kulit kertas projek  
*(as it appears on the title page and front cover of project paper)*

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.  
*(that the project paper acceptable in form and content, and that a satisfactory knowledge of the filed is covered by the project paper).*

Nama Penyelia Utama  
*(Name of Main Supervisor):* **MRS. SITI SAKIRA KAMARUDDIN**

Tandatangan  
*(Signature)* :  Tarikh (Date): 26/10/05

Nama Penyelia Kedua  
*(Name of 2<sup>nd</sup> Supervisor):* **MR. MOHD. SHAMRIE SAININ**

Tandatangan  
*(Signature)* :  Tarikh (Date): 26/10/05

## **PERMISSION TO USE**

In presenting this thesis in partial fulfillment of the requirements for the postgraduate degree from Universiti Utara Malaysia, I agree that University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Dean of the Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Request for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

**Dean of Faculty of Information Technology  
Universiti Utara Malaysia  
06010 UUM Sintok  
Kedah Darul Aman**

## **ABSTRAK**

Kewujudan pelbagai sumber maklumat dalam dunia maya serta korporat masa kini telah menarik minat para penyelidik dalam membantu melaksanakan proses pencarian, penapisan, dan pengurusan maklumat dengan lebih tepat lagi. Teknik pengkategorian teks merupakan salah satu teknik yang boleh diadaptasikan di dalam situasi ini. Kajian ini menerangkan mengenai teknik pengkategorian teks berdasarkan algoritma naive Bayes. Algoritma ini telah lama digunakan dalam tugas penkategorian teks. Pengkelasan naive Bayes adalah berdasarkan model kebarangkalian yang menggabungkan andaian bebas yang kukuh di mana ianya tiada kaitan dengan keadaan realiti. Tujuan kajian ini adalah untuk mengkategorikan dokumen yang berbentuk teks dengan menggunakan algoritma naive Bayes dan untuk mengukur sejauh mana teknik yang dipilih ini dapat mengkategorikan dokumen dengan betul. Kajian ini juga turut membincangkan mengenai eksperimen yang telah dijalankan dalam mengkategorikan artikel dengan menggunakan naive Bayes. Hasil daripada eksperimen ini menunjukkan ketepatan bagi 'training' adalah sebanyak 81.82% manakala ketepatan bagi 'testing' pula adalah sebanyak 47.62%.

## **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. The result shows that the accuracy for training is 81.82% whereas the accuracy for testing is 47.62%.

## **ACKNOWLEDGEMENT**

First of all I am so thankful to Allah for giving me the courage, ability, and strength to complete this project.

I would like to express my gratitude sincere appreciation to both of my supervisor Mrs. Siti Sakira bt Kamaruddin and Mr. Mohd Shamrie bin Sainin for guiding the research presented in this dissertation. They have been a constant source of motivation and encouragement. I appreciate and thank them for their continuous support, for being always accessible and for providing invaluable feedback on my work. Their encouragement helped shape the direction of my work. Without them, maybe this project is not completed properly.

I give my warm thanks to my friends for their support, encouragement, and collaboration in sharing ideas during this project. The persons that I mentioned above are Wan Faridah Hanum bt Wan Yaacob, Noraini bt Omar, and Hamidah bt Achmad. Thanks for being a wonderful person to me.

I would also like to thank to my beloved family especially my parents, Wan Ismail bin Wan Ali and Siti Hawa bt Mat Yusoff who never stop giving me a support, love, prayers, and always been there for me. Your love gives me strength to complete this study. Thanks to all my sisters and brothers for your continuous support.

Finally, I thank to everybody that involved in my project direct or indirectly for fruitful interactions and for their support.

## **TABLE OF CONTENTS**

PERMISSION TO USE	i
ABSTRAK	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	xi
LIST OF ABBREVIATIONS	xii

## **CHAPTER 1: INTRODUCTION**

1.1 Overview of Study	1
1.1.1 Naive Bayes	4
1.2 Problem Statement	6
1.3 Objectives of Study	6
1.4 Scope of Study	7
1.5 Significance of Study	8
1.6 Organization of the Report	8

## **CHAPTER 2: LITERATURE REVIEW**

2.1 Unstructured Data	11
2.2 Text Categorization	13



2.3 Machine Learning	16
2.4 Naive Bayes	21
2.5 Other Technique	27

## **CHAPTER 3: RESEARCH METHODOLOGY**

3.1 Introduction	34
3.2 Research Methodology of the Study	37

## **CHAPTER 4: PROTOTYPE DESIGN**

4.1 Rational Rose 2000	47
4.1.1 Unified Modeling Language (UML)	48
4.2 The General Architecture Design	48
4.3 The UML Diagram	50

## **CHAPTER 5: IMPLEMENTATION**

5.1 The Tool	60
5.1.1 Microsoft Visual Basic (VB) version 6.0	60
5.2 The Process and Algorithm for Text Categorization	60
5.3 The General Preview of the Text Categorization Prototype System	69

## **CHAPTER 6: RESULTS AND DISCUSSIONS**

6.1 Overview	76
6.2 Result	77

## **CHAPTER 7: CONCLUSION & RECOMMENDATION**

7.1 Conclusion	88
7.2 Limitation and Recommendation	90

## **REFERENCES**

REFERENCES	92
------------	----

## **APPENDIXES**

APPENDIX A	97
APPENDIX B	101

## LIST OF FIGURES

<b>Figure 1.1</b>	Definition of Machine Learning	3
<b>Figure 2.1</b>	Field of Study	17
<b>Figure 2.2</b>	Learning Algorithm for Categorization	19
<b>Figure 2.3</b>	Causal Structures of the Supervised Learning and Unsupervised Learning	21
<b>Figure 2.4</b>	Naive Bayes Classifier	23
<b>Figure 2.5</b>	Graphical Depiction of Naive Bayes Model	23
<b>Figure 2.6</b>	SVM Algorithm	28
<b>Figure 2.7</b>	k-NN Algorithm	29
<b>Figure 2.8</b>	Decision Tree Algorithm	30
<b>Figure 2.9</b>	KDD Process of Unstructured Data	33
<b>Figure 3.1</b>	Outputs of Design Research	36
<b>Figure 3.2</b>	The General Methodology of Design Research	37
<b>Figure 3.3</b>	Example of Article Classification	45
<b>Figure 4.1</b>	The Architecture Design for Text Categorization Prototype System	49
<b>Figure 4.2</b>	Use Case Diagram	51
<b>Figure 4.3</b>	Class Diagram	53

<b>Figure 4.4</b>	Sequence Diagram for Preprocess	55
<b>Figure 4.5</b>	Sequence Diagram for Learning	56
<b>Figure 4.6</b>	Sequence Diagram for Classify	57
<b>Figure 4.7</b>	Sequence Diagram for Testing	58
<b>Figure 5.1</b>	Text Preprocessing	62
<b>Figure 5.2</b>	Algorithm for Preprocess	62
<b>Figure 5.3</b>	tblAbstract	63
<b>Figure 5.4</b>	Code for Preprocess	63
<b>Figure 5.5</b>	Algorithm for Learning	64
<b>Figure 5.6</b>	tblClassProb	65
<b>Figure 5.7</b>	tblVocabulary	66
<b>Figure 5.8</b>	Code to Calculate Probability Category	66
<b>Figure 5.9</b>	Code to Calculate Word Probability	67
<b>Figure 5.10</b>	Algorithm for Classification	68
<b>Figure 5.11</b>	tblAbstractTest	68
<b>Figure 5.12</b>	Algorithm for Preprocess	69
<b>Figure 5.13</b>	Main Page for Text Categorization Prototype System	70
<b>Figure 5.14</b>	The Result's for Preprocess Function	71
<b>Figure 5.15</b>	The Result's for Learn Function	72
<b>Figure 5.16</b>	The Result's for Classify Function	73

<b>Figure 5.17</b>	The Result's for Test Function	74
<b>Figure 6.1</b>	The Percentage of Correctly Classified Articles for Each Category (Training)	80
<b>Figure 6.2</b>	The Percentage of Correctly Classified Articles for Each Category (Testing)	85
<b>Figure 6.3</b>	The Comparison between Training and Testing	87

## LIST OF TABLES

<b>Table 2.1</b>	Comparison of Naive Bayes and Adaptive Bayes Network	25
<b>Table 4.1</b>	Use Case Description	51
<b>Table 6.1</b>	Result for Training	78
<b>Table 6.2</b>	Summarization of Classification (Training)	79
<b>Table 6.3</b>	Confusion Matrix for Training Articles	81
<b>Table 6.4</b>	Result for Testing	83
<b>Table 6.5</b>	Summarization of Classification (Testing)	84
<b>Table 6.6</b>	Confusion Matrix for Testing Articles	85

## LIST OF ABBREVIATIONS

<b>AI</b>	Artificial Intelligence
<b>BASIC</b>	Beginners' All-Purpose Symbolic Instruction Code
<b>BLOB</b>	Binary Large Objects
<b>GIS</b>	Generalize Instance Set
<b>GUI</b>	Graphic User Interface
<b>IDF</b>	Inverse Document Frequency
<b>IS</b>	Information System
<b>IT</b>	Information Technology
<b>KDD</b>	Knowledge Discovery in Database
<b>KM</b>	Knowledge Management
<b>KMICE</b>	Knowledge Management International Conference
<b>k-NN</b>	k-Nearest Neighbor
<b>OOP</b>	Object-Oriented Programming
<b>RAD</b>	Rapid Application Development
<b>RAM</b>	Random Access Memory
<b>SOM</b>	Self-Organizing Map
<b>SVM</b>	Support Vector Machine
<b>TF</b>	Term Frequency
<b>UML</b>	Unified Modeling Language
<b>VB</b>	Visual Basic

# **CHAPTER ONE**

## **INTRODUCTION**

This chapter presents the main idea of this study which is text categorization process. It provides an overview of the technique that was used to categorize articles that is naive Bayes. Naive Bayes is one of the learning algorithms for machine learning. In addition, this chapter also discusses problem statement, objectives, scope, and the significance of the study.

### **1.1 OVERVIEW OF STUDY**

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.



The contents of  
the thesis is for  
internal user  
only

## 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](http://en.wikipedia.org/wiki/Bayesian_probability).
- Basic ODM concept (2002). Oracle Corporation. Retrieved June 26, 2005, from <http://www.cs.umb.edu/cs634/ora9idocs/datamine.920/a95961/1concept.htm>.
- Basu, A., Watters, C., & Shepherd, M. (2002). Support vector machines for text categorization. *Proceedings of the 36<sup>th</sup> Hawaii International Conference on System Sciences (HICSS'03)*. Retrieved July 6, 2005, from IEEE Xplorer database.
- Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2001). On feature distributional clustering for text categorization. Retrieved June 23, 2005, from ACM Digital Library database.
- Bennet, S., McRobb, S., & Farmer, R. (2002). *Object oriented system analysis and design using UML* (2<sup>nd</sup> ed.) McGraw-Hill Education: Backshire.
- Bengio, S. (2005). Statistical machine learning. Retrieved August 14, 2005, from <http://www.idiap.ch/~bengio/lectures/intro.pdf>.
- Booch, G., Jacobson, I., & Rumbaugh, J. (2001). *The unified modeling language user guide*. Addison-Wesley: Boston.
- Categorization. (2005, August 29). Wikipedia Encyclopedia. Retrieved June 4, 2005, from <http://en.wikipedia.org/wiki/Categorization>.
- Caragea, D. (2004). Learning Classifier from distributed, semantically heterogeneous, autonomous data sources. Retrieved August 18, 2005, from <http://www.cs.iastate.edu/~honavar/Papers/caragea-thesis.pdf>.
- Chaovalit, P., & Zhou, L. (2005). Movie review mining: a comparison between supervised and unsupervised classification approaches. *Proceedings of the 38<sup>th</sup> 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.

- Christiani, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge University Press.
- Confusion matrix. (2005, August 27). Wikipedia Encyclopedia. Retrieved July 28, 2005, from [http://en.wikipedia.org/wiki/Confusion\\_matrix](http://en.wikipedia.org/wiki/Confusion_matrix).
- 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. (1998). Inductive learning algorithms and representations for text categorization. Retrieved July 23, 2005, from <http://research.microsoft.com/~sdumais/cikm98.pdf>.
- Gurney, K. (n.d). Supervised learning. Retrieved August 5, 2005, from [http://www.shef.ac.uk/psychology/gurney/notes/l10/subsubsection3\\_3\\_6\\_2.html](http://www.shef.ac.uk/psychology/gurney/notes/l10/subsubsection3_3_6_2.html)
- Gurusamy, S., Manjula, D., & Geetha, T., V. (2002). Text mining in 'request for comments documents series'. *Proceedings of the Language Engineering Conference*. Retrived June 20, 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.
- Iiritano, S., & Ruffalo, M. (2001). Managing the knowledge contained in electronic documents: a clustering method for text mining. Retrived June 20, 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.
- Iwayama, M., & Tokunaga, T. (1995). Cluster-based text categorization: a comparison of category search strategies. Retrieved August 19, 2005, from ACM Digital Library database.
- Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The Unified Software Development Process*, Reading, MA: Addison-Wesley, ACM Press.
- Jian-Hong, L., & Tsui-Feng, H. (2004). Fuzzy correlation and support vector learning approach to multi-categorization of documents. *IEEE International Conference on Systems, Man and Cybermetics*. Retrieved July 7, 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](http://www.cs.cornell.edu/People/tj/publications/joachims_98a.pdf)
- Jun, H., & Houkuan, H. (2002). An algorithm for text categorization with SVM. *Proceedings of IEEE TENCON'02*. Retrieved July 7, 2005, from IEEE Xplorer database.
- Kao, A., Quach, L., Poteet, S., & Woods, S. (2003). User assisted text classification and knowledge management. Retrived June 5, 2005, from ACM Digital Library database.
- Kantarcioğlu, M., & Vaidya, J. (2003). Privacy preserving naive bayes classifier for horizontally partitioned data. Retrieved August 7, 2005, from <http://www.cis.syr.edu/~wedu/ppdm2003/papers/1.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.
- Kuhn, T. (1996). The structure of scientific revolutions. Chicago, University of Chicago Press.
- Lam, W., Member, IEEE, & Han, Y. (2003). Automatic textual document categorization based on generalized instance sets and a metamodel. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5). Retrieved July 6, 2005, from IEEE Xplorer 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.
- March, S. & Smith, G. (1995). Design and Natural Science Research on Information Technology. *Decision Support Systems* 15.
- 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](http://en.wikipedia.org/wiki/Naive_Bayes).
- Naive Bayes rule generator. (2005). Retrieved June 4, 2005, from [http://grb.mnsu.edu/grbts/doc/manual/Naive\\_Bayes.html](http://grb.mnsu.edu/grbts/doc/manual/Naive_Bayes.html)
- 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>.

- Pavlov, D., Balasubramanyan, R., Dom, B., Kapur, S., Parikh, J. (2004). Document preprocessing for naive bayes classification and clustering with mixture of multinomials. Retrieved June 29, 2005, from ACM Digital Library database.
- 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., Bouzerdoun, 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.
- Purao, S. (2002). Design research in the technology of information systems: truth or dare. GSU Department of CIS Working Paper. Atlanta.
- 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>.
- Rao, R. (2003). From unstructured data to actionable intelligence. *IEEE Computer Society*. Retrived June 20, 2005, from IEEE Xplorer database.
- 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>.
- Rossi, M., & Sein, M. (2003). Design Research Workshop: A Proactive Research Approach. *IRIS 26, August 9 – 12, 2003*. Retrieved August 15, 2005, from [http://tiesrv.hkkk.fi/iris26/presentation/workshop\\_designRes.pdf](http://tiesrv.hkkk.fi/iris26/presentation/workshop_designRes.pdf).
- 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.pdf](http://www.cs.princeton.edu/courses/archive/spring03/cs511/scribe_notes/0204.pdf)
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Survey* 34(1). Retrived June 15, 2005, from ACM Digital Library database.
- 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>.
- Simon, H. (1996). *The Sciences of the Artificial*, Third Edition. Cambridge, MA, MIT Press.

- The Rose JADE Link. (2003). Jade Software Software Corporation. Retrieved August 26, 2005, from <http://www.jadeworld.com/downloads/jade6/RoseJADELLink.pdf>.
- Uimonen, T. (2000). Case: rational rose. 81940 A Seminar on Reverse Engineering. Retrieved August 26, 2005, from <http://www.cs.tut.fi/~tsysta/sem/Reports/Rose.pdf>.
- Valpola, H. (2000, Oct 31). Supervised vs. unsupervised learning. Retrieved August 8, 2005, from [http://www.cis.hut.fi/harri/thesis/valpola\\_thesis/node34.html](http://www.cis.hut.fi/harri/thesis/valpola_thesis/node34.html).
- Vinciarelli, A. (2004). Noisy text categorization. *Proceeding of the 17th International Conference on Pattern Recognition (ICPR'04)*. Retrieved August 12, 2005, from IEEE Xplorer database.
- Visual basic overview. (2002). Information Technology Toolbox. Retrieved August 26, 2005, from <http://visualbasic.ittoolbox.com/browse.asp?c=VBPeerPublishing&r=%2Fpub%2Fvb%5Foverview%2Ehtm>
- 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.
- Yu, F., An, J., Li, H., Zhu, M., & Yang, O. (2004). Intelligence text categorization based on Bayes algorithm. *Proceedings of 2004 International Conference on Information Acquisition*. Retrieved July 6, 2005, from IEEE Xplorer database.
- Zadok, E. (2001). Naive Bayes. Retrieved August 1, 2005, from <http://www.fsl.cs.sunysb.edu/docs/binaryeval/node5.html>.