SPAM FILTERING USING BAYESIAN TECHNIQUE BASED ON INDEPENDENT FEATURE SELECTION

MASURAH BINTI MOHAMAD

A project report submitted in partial fulfillment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

APRIL, 2006

“*To my beloved family, thanks for your support and sacrifice. To all my friends, nice knowing you all and thanks for the understanding and encouragement.*”

ACKNOWLEDGEMENTS

Alhamdulillah, it is with Allah S.W.T will that I get to finish this project in time given. Here, I would like to take this opportunity to thank my supervisor, Dr. Ali bin Selamat for his attention and guidance throughout the length of this study. Without his help, I would be lost and knowing nothing. Not forgetting, my thanks go to Associate Professor Dr. Siti Mariyam Binti Shamsuddin for her helps and suggestions in making this project more interesting.

Special thanks to all my course mates, friends, staff, and lecturers in the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia for their help and support. Not forgetting, to my family for their supports and understanding.

ABSTRACT

Bayesian technique is one of the classification techniques which can be applied to a certain problem domain such as classification task. Therefore, this technique had been chosen to conduct a classification task with emails dataset where the emails are comprised of spam and non spam emails. Bayesian technique has been applied to observe whether it can produce a good result in spam emails classification or not. Beside, this project also applied Rough set as a comparison technique to classify the spam emails. The classification task is done based on the independent feature selection where only one most occurrence term for each email is chosen as an input to the Bayesian probability. Some of the measurement evaluation had been used to evaluate the classification performance. The measurements are precision, recall, sensitivity, specificity, accuracy and error rate. After the measurements process, these two technique were compared to identify which one of these two techniques is best in classifies spam emails based on the experimental results. The results show that Bayesian technique is good than Rough set technique in classifies spam emails. However the results also indicate that Rough set also suitable for spam filtering problem. Finally, some suggestions were being discussed so that this project can be improved in future work to get a better result compared to the current result which had been retrieved in this project.

ABSTRAK

Teknik Bayesian adalah salah satu daripada teknik pengkelasan yang sering digunakan untuk menyelesaikan sesuatu domain masalah. Teknik ini telah dipilih untuk melaksanakan proses pengkelasan yang melibatkan set data yang terdiri daripada emel iaitu emel *spam* dan emel *non spam*. Teknik ini telah dilaksanakan dengan jangkaan untuk melihat keberkesanannya dalam menjalankan proses pengkelasan emel sama ada emel tersebut adalah *spam* atau *non spam*. Selain itu, projek ini juga menggunakan Rough set sebagai teknik perbandingan. Pemilihan ciri yang *independent* telah dipilih sebagai metod utama iaitu menggunakan satu sahaja perkataan yang paling banyak muncul bagi setiap emel sebagai input kepada proses pengelasan. Beberapa pengukuran penilaian iaitu *precision*, *recall*, *accuracy*, *specificity*, *sensitivity* dan *error rate* turut dilaksanakan untuk menilai prestasi kedua-dua teknik ini selepas melaksanakan proses pengkelasan. Setelah proses pengukuran penilaian dilaksanakan, keputusan-keputusan daripada penilaian tersebut akan dianalisa untuk membandingkan teknik yang manakah adalah yang terbaik dalam melaksanakan proses pengkelasan emel *spam*. Keputusan eksperimen yang diperolehi, menunjukkan bahawa Bayesian lebih tepat dalam mengelaskan emel spam berbanding teknik Rough set. Walaubagaimanapun dapat disimpulkan bahawa teknik Rough set juga amat sesuai diaplikasikan dalam proses pengkelasan emel spam. Berdasarkan keputusan yang diperolehi beberapa cadangan pembaikan juga dinyatakan dengan harapan projek ini akan diperbaiki dan dipertingkatkan lagi keputusan pengkelasan supaya lebih berkesan daripada yang diperolehi sekarang.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\overline{B}X$	Upper approximation
$\underline{B}X$	Lower approximation
n_i	Number of occurrences on considered term
X	Set of
Y	Negative region
$\sum_k n_k$	Number of occurrences of all term

LIST OF ABBREVIATION

ASSP	Anti Spam SMTP Proxy
DCC	Distributed Checksum Clearinghouse
DF	Document Frequency
ICA	Independent Component Analysis
IDF	Inverse Document Frequency
ISP	Internet Server Protocol
MTA	Message Transfer Agent
PCA	Principle Component Analysis
RBL	Realtime Black Hole List
SVM	Support Vector Machine
TDMA	Tagged Message DeliveryAgent
TF	Term Frequency
TFIDF	Term Frequency Inverse Document Frequency

CHAPTER 1

INTRODUCTION

Spam is also known as unsolicited commercial email. The other popular name is junk mail which floods the Internet users' electronic mailboxes. These junk mails can contain many types of messages such as pornography, commercial advertising, get rich quick scheme, doubtful product, viruses or quasi legal services. There are two types of spam which are Usenet spam and direct mail messages. Usenet spam is a single message that is sent to 20 or more Usenet newsgroups which will target the people who read newsgroups but rarely or never give their email addresses away (Mueller, 1999). Spam that is categorized as a direct mail message is when it targets individual emails messages by stealing the Internet mailing list, scanning the Usenet postings or searching the web for addresses. Email spamming will get worse if the recipient reply to the messages, which will cause the recipients' addresses available to be attacked by other spammers.

Some email users are busy to open their mailboxes until it become crowded with unwanted emails and make the received emails being left unfiltered. Recipients of spam should spend a lot of their time to delete and classify which of those emails are spam or legitimate. When a user is flooded with a large amount of spam, the chance of he or she forgot to read a legitimate message increases. As a result, many email readers will have to spend a non-trivial portion of their time removing unwanted messages. Spam also creates a burden on mail servers and Internet traffic, all for unwanted messages (Stone, 2003).

Spam filtering is a kind of text classification task. Spam filtering techniques can be classified into two categories which are specific and general (Westbrook, 2000a). Specific filtering technique looks at the characteristics from actual spam messages, such as key phrases or words, source of the spam, or specific action requested of the recipient (e.g. Bayesian filtering approach). While general filtering technique typically look for a series of suspicious elements in an e-mail like an all capitalized subject field, exclamation points in the subject field, or obvious phony names in the "From" (sender) field including numbers and other features. However, specific filtering technique is more accurate than general filtering technique because it requires more maintenance by the vendor and especially from the mail user themselves because they depend on someone to set or classify the actual spam message to create and update the spam signatures.

There are two parameters which consist of effectiveness and accuracy used to measured spam filtering systems (Westbrook, 2000b). Effectiveness is measured by the percentage of spam that is caught. This percentage should be as high as possible. Accuracy is measured by the percentage of e-mails incorrectly identified as spam. This second percentage should be as low as possible. A well-designed of spam filters can score well on both measures.

Several different approaches have been examined to filter the spam; many of those approaches are widely used. One broad approach that attacks spam at the network level is Rough set technique (Wang et al, 1999). Rough set based filter can be used to detect junk emails on the Internet. One major technique of it is to build filters in email transfer route (Wang et al, 1999). For instance, some servers block spam based on so-called Realtime Black Hole lists (RBLs), which alert servers of undergoing flooding spam. RBLs are the aggressive bouncers of the Internet; anything listed in the RBL will be rejected (Schroder, 2003). Another broad approach examines the content of an incoming message for features which indicate its status as spam or legitimate. In present, researchers are interested to apply statistical learning algorithm such as Bayesian introduced by Thomas Bayes (http://www.wikipedia.org), support vector machine technique (SVM) suggested by Corinna Cortes and Vladimir Varnik in 1995 (http://www.wikipedia.org) and AdaBoost algorithm, formulated by Freund and Schapire (http://www.wikipedia.org). Taking the latter approach and treating e-mail filtering as a text classification problem, researchers have applied several statistical learning algorithms to email dataset with promising results including problems more difficult than spam filtering, such as author identification (Oliver et al, 2000). These approaches were applied in order to produce good results in filtering spam

1.1 Background of Problem

In order to avoid the growing problem of junk E-mail on the Internet, several methods have been examined for the automated spam filtering system which is responsible in eliminating such unwanted messages from a recipient electronic mailbox. Regarding to this problem, it has become necessary for us to have a filtering system that will classify the emails either as junks or legitimate. Junk mails are the unwanted messages which occurred without the permission of the recipients such as pornographic message. Meanwhile, legitimate emails are messages which occurred with the permission of the recipients such as the messages from family and friends. Many commercial products are available nowadays which allow users to create a set of logical rules to filter junk mails. For example, there is a system that requires users to create a rule set to detect junk or spam assuming that their users are knowledgeable to create robust rules. This solution, however, is quite difficult and fussy to the busy and limited (have not much knowledge about spam or even a computer) users. Moreover, as the nature of junk mail changes over time, these rule sets must be constantly tuned and refined by the user. This is a time-consuming and often tiresome process which can be closely generating an error during system execution and not generating the spam filter.

Filtering system must have a capability to independently detect the characteristics of spam. In other word, a junk mail filtering system should be able to automatically adapt to the changes in the characteristics of junk mail over time (Sahami et al, 1998). Moreover, by having a system that can learn directly from data in a user's mail repository, such junk filters can be personalized to the particular characteristics of a user's legitimate (and junk) mail. For example some of the users especially the companies that are involved in mortgage business such as loan brokers and correspondent lenders will classify the ‘mortgage’ phrase that contain in their emails as a legitimate and other user that assume ‘mortgage’ phrase is not important and should not be in their inboxes will delete or classify the phrase as junk. Hence, this particular task will easily lead to the development of much more accurate junk filters for each user. Along these lines, methods have recently been suggested for automatically learning rules to classify email. While such approaches have shown some success for general classification tasks based on the text of messages, they have not been employed specifically with the task of filtering junk mail in mind. As a result, such systems are not focused on the specific features which distinguish junk from legitimate email (Cohen, 1996). Researchers still are doing great study in fighting back the spam by developing or producing such a good filtering system and algorithm that will detect spam at most efficient ways.

Researchers have applied a variety of statistical learning methods in automatically generating probabilistic email text classification models such as the Naïve Bayesian classifier (Lewis & Ringuette, 1994) (Mitchell, 1997) (McCallum et al, 1998), support vector machine (SVM) technique (Drucker et al, 1999), genetic programming (Katiraj, 1999), decision tree classifier (Diao et al, 2000) as well as more expressive Bayesian classifiers (Koller & Sahami, 1997).

From all the studies done by those researches in the past, Bayesian classification technique has been favored in the area of email filtering compared to Rough set technique because of its precision and efficiency. Recently, Bayesian classification techniques have been applied in solving problems of spam filtering. In this project, a comparison between Bayesian and Rough set filtering techniques will be performed to discover which of these techniques can prove the effectiveness in producing good filter results in the domain problem. It is essential to compare between the classifier in different domain problems rather than employing only one method to discover the best result. The study of comparing the various techniques in spam filtering will definitely do a benefit towards the research in this area.

1.2 Problem Statement

1.2.1 Unsolicited commercial emails are extremely flooding our electronic mailboxes.

1.2.2 Spammers have many ways or techniques to flood the users' emails repository such as by obtaining email addresses from Usenet postings, DNS listings, Web pages, guessing common names at known domains and searching for email addresses corresponding to specific users.

1.2.3 It is quite difficult and wasting of time to fight these spam emails without having such good filtering systems or tools that help to classify the incoming emails.

1.2.4 This comparative study is done in order to determine whether Bayesian and Rough set techniques are able to filter the spam.

1.2.5 Moreover this project also tries to determine which of these techniques give better results in spam filtering.

1.3 Objectives of the Thesis

1.3.1 To apply Bayesian and Rough set techniques in filtering spam.

1.3.2 To compare the effectiveness of the filtering results from Bayesian and Rough set techniques.

1.3.3 To find out whether Bayesian and Rough set techniques can be apply on emails dataset in order to perform a better result in classification.

1.4 Scope

1.4.1 Use 450 lists of e-mails or data sets as training data.

1.4.2 Microsoft Outlook has been selected to be the email platform.

1.4.3 Employ precision and recall concepts to test the effectiveness of the Bayesian classification results.

1.4.4 Apply filtering technique on emails with text only.

1.4.5 Total emails for training process are 300 and for testing process are 150 emails.

1.4.6 Stemming process is done on Malay and English languages.

1.4.7 The classification process only focuses on subject and body of the messages.

1.4.8 Apply only one feature selection or one attribute which is the term that mostly occurs in each email.

1.4.9 Apply ROSETTA toolkit software to classify emails for Rough Set technique.

1.4.10 Use my own personal emails instead of using benchmark emails (Hovold, 2005).

1.5 Project Plan

This project is carried out in two semesters. The first part of the project focuses on understanding the general view of spam problem and filtering technique and also the past approaches that have been applied by other researches as well as methodology to be used in this project. Most of the time in the first semester is used to explore and gather relevant information from the text books and published journals. The total understanding in spam filtering and artificial intelligence methods is important in order to know the different methods that can be used in solving filtering problems. At the end of first semester, a better understanding of Bayesian and Rough set spam filtering techniques is achieved and preliminary results will be produced to show that this comparative study can be proceed or not before executing the rest of the complicated works. The report for the first semester includes Introduction, Literature View, Methodology and Initial Findings.

The second part of the project involves implementing Bayesian and Rough set filtering techniques for email classification whether the email is junk or legitimate. The implementation will begin with preprocessing works includes feature extraction (stemming and stop word removal) and feature selection such as term frequency-inverse document frequency (*tfidf*), and then apply probabilistic classification to classify emails. Next, Rough set technique is applied to compare the effectiveness of spam filtering between itself and the Bayesian technique. Comparison will be carried out based on classification and the efficacy of Bayesian and Rough set filtering technique. The second part of the report includes Experimental Result and Conclusion.

1.6 Thesis Contribution

This project will give better insights and idea or solution in the use of email filtering to classify whether the emails are junk or legitimate. This comparative study also may suggest which of the filtering techniques should be used to achieve the effectiveness of spam filtering; Bayesian or Rough set techniques.

1.7 Conclusion

Nowadays, email becomes one of the most important communication tools for people around the world. Through email, we can communicate with other people easily, faster than sending a letter and of course the service is for free. However, the email service also has some problem to overcome such as viruses and spam flooding especially. There are many researches and works that had been done to defeat this spam problem and mostly were successful in filtering the spam. Therefore, this project was being suggested in order to fulfill the objectives which are to apply, to test and to compare between these two filtering techniques; Bayesian and Rough set either one of these technique gives a good result in filtering the spam.

REFERENCES

Abraham P.M. (2003). "*Black Hole List*". Referred on 17th October 2005 from World Wide Web: (http://www.dynamicnet.net/news/articles/move_to_rbl.html)

Ahmed S. and Mithun F. (2004). "*Word Stemming to Enhance Spam Filtering*". In Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

Bong Mei Yen (2005). *Pengenalpastian Topik Perbincangan Bagi Sesuatu Ruang Perbualan Atas Talian Menggunakan Teknik Principal Component Analysis*. Universiti Teknologi Malaysia: Projek Sarjana Muda.

Bill B.W., McKay R.I., Abbas A.H. and Barlow M. (2000). "*A Comparative Study for Domain Ontology Guided feature Extraction*". Proceedings of the twenty-sixth Australasian computer science conference on Conference in research and practice in information technology, p.69-78, February 01, 2003, Adelaide, Australia

Bingham E., Kuusisto J. and Lagus K. (2002). "*ICA and SOM in Text Document Analysis*". The 25th ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, pp. 361-362.

Breault, Joseph L. (2001), "*Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?*". Computing Science and Statistics, 33, /I2001Proceedings/JBreault/JBreault.pdf

Cranor L.F. and Brian A.L. (1998). "*Spam!*". Communications of the ACM, Vol.41, No.8, pp. 74-83, Aug. 1998.

Chouchoulas A. and Shen Q (1999). "*A Rough Set-Based Approach to Text Classification*". Proceedings of the 7th International Workshop on Rough Sets (Lecture Notes in Artificial Intelligence, No. 1711), pages 118-127, 1999.

Cohen W. W. (1996). "Learning Rules that Classify E-Mail" AAAI Spring Symposium on ML and IR 1996.

Cunningham, Padraig (2004). "*Dimension Reduction and Feature Subset Selection*". Presentations MUSCLE Scientific Meeting Malaga, 4-5 Nov 2004.

David D. Lewis (2004). "*(Naïve) Bayesian Text Classification for Spam Filtering*". In Proceedings of ASA Chicago Chapter Spring Conference, Loyola University, 2004.

Diao Y., H. Lu, and D. Wu (2000). "*A comparative study of classification-based personal e-mail filtering*". In Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'00), pages 408--419, Kyoto, JP.

Drucker, H., Wu, D., Joksons, D.W. (1999). "*Support Vector Machine for spam categorization*". IEEE Trans on Neural Networks, 10:1048-1054.

Fuka K. and Hanka Rudolf (2001). "*Feature Set Reduction for Document Classification Problems*". IJCAI-01 Workshop: Text Learning: Beyond Supervision. Seattle (August 2001).

Frasconi P. Soda G. and Vullo A.(2001). "*Text Categorization for Multiple page Documents: A Hybrid Naïve-Bayes HMM Approach*". ACM-IEEE Joint Conference on Digital Libraries, 2001.

Garg A. and Roth D. (2001). "*Understanding Probabilistic Classifiers*". Conference Proceeding. ECML'01, Sept. 2001.

Gaustad T. and Bouma G. (2002). "*Accurate Stemming of Dutch for text Classification*".
In Mariët Theune and Anton Nijholt, editors, Computational Linguistics in the Netherlands (CLIN) 2001, Twente University, 2002.

Graham P. (2002). "*A Plan for Spam*".
Referred on 19th October 2005 from World Wide Web:
(www.paulgraham.com/spam.html)

Graham P. (2003). "*Better Bayesian filtering*". In Proceeding of the First Annual Spam Conference. MIT. (www.paulgraham.com/better.html)

Gregory L.W. And Wu S.F. (2004). "*On Attacking Statistical Spam Filters*". In Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

Grenager H. S. (1996). "*Rough Sets*".
Referred on 19th October 2005 from World Wide Web:
http://www.pvv.ntnu.no/~hgs/project/report/node38.html

Hastings S. (2003). "*Fight Spam with SpamProbe*".
Referred on 17th October 2005 from World Wide Web:
www.linuxjournal.com.

Holden S. (2002). "*Spam Filtering II*".
Referred on 17th October 2005 from World Wide Web:
(http://sam.holden.id.au)

Hovold, Johan (2004). "*Naive Bayes Spam Filtering Using Word-Position-Based Attributes*". Lund University, Lund, Sweden.

Hui, Shirley (2002). "*Rough Set Classification of Gene Expression Data*". CS 798 Final Project. University of Waterloo.

Jason D.M. Rennie and Rifkin R. (2002). "*Improving Multiclass text Classification with the SVM*". AI Memo, AIM-2001-026, MIT, 2001.

Jiye Li and Nick Cercone (2005). "*Empirical Analysis on the Geriatric Care Data Set Using Rough Sets Theory*". Technical Report, CS-2005-05, School of Computer Science, University of Waterloo.

Joachims T. "*Text Categorization with Support Vector Machines: Learning with Many Relevant Features".* Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

John G., Kohani R. and Pfleger K. (1994). "*Irrelevant Features and The Subset Selection Problem*". To appear. Honavar, V. and Uhr, L. (1993). Generative learning structures for generalized connectionist networks. Information Sciences, 70(1-2):75-108

Judge P. (2005). "*How Spammers Fool Spam Filters*"
Referred on 18th October 2005 from World Wide Web:
(http://www.securitydocs.com/library/3436)

Kiritchenko K. and Matwin S. (2001). "*E-Mail Classification with Co-Training*".
Referred on 18th October 2005 from World Wide Web:
(www.site.uottawa.ca/~stan/papers/2001/cascon2002.pdf)

Koller D. and Sahami M. "*Hierarchically Classifying Documents Using Very Few Words*". International Conference on Machine Learning (ICML)1997: 170-178.

Korfhage R.R. (1997). "*Information Storage and Retrieval*". John Wiley & Sons, Inc., New York, NY, 1997.

Lewis D.D. and Ringuette M. (1995). *"Comparison of Two Learning Algorithms for text Categorization"*. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR' 94).

Lina Z. and Jiao X. (2004). "*A Collaborative Approach to Spam E-Mail Filtering: Recommendation Using Ontological User Profiling*".

Luz S. (2005). "*Text Classifier Induction: Decision Trees*".
Referred on 17th October 2005 from World Wide Web:
https://www.cs.tcd.ie/courses/ baict/baim/ss/part2/dtreetc.pdf

Magnani M. (2003). "*Technical Report on Rough Set Theory for Knowledge Discovery in Data Bases*".
Referred on 17th October 2005 from World Wide Web:
(magnanim.web.cs.unibo.it/data/pdf/roughkdd.pdf)

Marasek K. (1997). "Methods of Data Classification". Experimental Phonetic Group. Institute of Natural Language Processing University of Stuttgart, Germany.
Referred on 17th October 2005 from World Wide Web:
Http://www.ims.uni-stuttgart.dc/phonetic/EGG/pagev4.htm

Mariah Binti Mohd Daud (2004). "*Pengelasan Email Mengikut Kategori Menggunakan Support Vector Machine (SVM)"*. Universiti Teknologi Malaysia: Projek Sarjana Muda.

Mertz D. (2003a). "*Spam Filtering Techniques: Six approaches to eliminating unwanted e-mail*".
Referred on 16th October 2005 from World Wide Web:
(http://www.opensourcetutorials.com/tutorials/Server-Side-Coding/Administration/spam-filtering-techniques/page6.html)

Mertz D. (2004b). "*Spam Filtering techniques, Rule Based Ranking*". Referred on 16th October 2005 from World Wide Web: (http://www.opensourcetutorials.com/tutorials/Server-Side-Coding/Administration/spam-filtering-techniques/page6.html)

Michael J. Frombeger (2004). "*Bayesian Classification of Unsolicited E-Mail*". Referred on 20th January 2006 from World Wide Web: (http://www.cs.dartmouth.edu/~sting/acad.shtml)

Mueller S.H. (1999). "*What is Spam*". Referred on 16th October 2005 from World Wide Web: (http://spam.abuse.net/overview)

Nicholas T. (2003). "*Using AdaBoost and Decision Stumps to Identify Spam E-Mail*". Referred on 18th October 2005 from World Wide Web: (nlp.stanford.edu/courses/ cs224n/2003/fp/tyronen/report.pdf)

O' Brien C. and Vogel C. (2002) *"Spam Filter: Bayes vs. Chi-squared; Letters vs. Words"*. Unpublished paper.

Orhn, A. and J. Komorowski (1997). "*ROSETTA: A Roughset Toolkit for Analysis of Data*". Joint Conference of Information Sciences: semiotics, fuzzy logic, soft computing, computer vision, neural computing, genetic algorithm, pattern recognition, evolutionary computing, Durham NC, Duke University Press: 403-407.

Painter James (2003). *"Uses of Bayesian Statistics"*. Referred on 12th October 2005 from World Wide Web: http://www.tessella.com/Literature/Supplements/PDF/bayesianstats.pdf.

Pawlak Zdzislaw (1982). "*Rough Sets*". International Journal of Computer and Information Sciences 11 (1982): 341-356.

Pearl Judea (1988). "Probabilistic Reasoning in Intelligent Systems". San Mateo, CA: Morgan Kaufman Publishers.

Pechenizkiy M, Puuronen S. and Tsymbal A. (2000). "*Feature Extraction for Classification in Knowledge Discovery Systems*". In: V.Palade, R.J.Howlett, L.C.Jain (Eds.), Proc. 7th International Conference on Knowledge-Based Intelligent Information & Engineering Systems KES'2003, Lecture Notes in Artificial Intelligence, Vol.2773, Heidelberg: Springer-Verlag, pp. 526-532.

Popovic M. and Willett P. (1992). "The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data". Journal of the American Society for Information Science (JASIS), Volume 43(5): 384-390 (1992).

Provost J. (2002). "*Naïve-Bayes vs Rule-Learning in Classification of Email*". The University of Texas at Austin, Artificial Intelligent Lab. Technical report AI-TR-99-284.

Rachel (2005). "*GFI Whitepaper – Why Bayesian Filtering*".
Referred on 19th October 2005 from World Wide Web:
(www.gfi.com)

Ramos J. (2001). "*Using TF-IDF to determine Word relevance in Document Queries*".
Referred on 17th October 2005 from World Wide Web:
(www.cs.rutgers.edu/~mlittman/ courses/ml03/iCML03/papers/ramos.pdf).

Sahami M., Susan Dumais, David Heckerman and Eric Horvitz (1998). "*A Bayesian Approach to Filtering Junk E-Mail*". In Learning for Text Categorization. Papers from the 1998 Workshop, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.

Sahami M., Dumais S., Platt J. and Heckerman D. (1998). "*Inductive Learning Algorithms and Representations for Text Categorization*". In Proceedings of ACM-CIKM98, Nov. 1998, pp. 148-155.

Schroder C. (2003). "*Real time Black-hole Lists: Heroic Spam Fighters or Crazed Vigilantes*".
Referred on 17th October 2005 from World Wide Web:
(http://www.enterprisenetworkingplanet.com/netsysm/article.php/1594561).

Selamat A. and Omatu S. (2003). "*Web Feature Selection and Classification using Neural Networks*". Information Sciences. Article in Press. January 2004. Volume 158: 69-88.

Steppe J.M. (1994). "*Feature and Model Selection in Feed Forward Neural Networks*". PhD Dissertation at Air Force Institute of Technology, Ohio

Stone T. (2003). "*Parameterization of naïve Bayes Spam Filtering*". Masters Comprehensive Exam, University of Colorado at Boulder, 2003.

Strackeljan J. (1999) "*Feature Selection Methods for Soft Computing Classification*". Proceedings of the ESIT 1999 (European Symposium on Intelligent Techniques), Kreta, Juni 1999.

Swiniarski R.W. and Skowron A. (2003). "*Roughset Methods Feature Selection and Recognition*". International Journal Application Math. Computer. Science., 2001, Vol.11, No.3, 565-582.

Swiniarski R. W. (2001). "*Rough sets Methods in Feature Reduction and Classification*". International Journal Application Math. Computer Science, Vol. 11, No3, 565-582.

Toma I. (2004). "*Optimization Techniques for Neural Networks Text Classifiers*". Next Web Generation Seminar.Summer 2004.

Tschabitscher H. (2005). "*What You Need to Know About Bayesian Spam Filtering*".
Referred on 16th October 2005 from World Wide Web:
(About.com. http://www.email.about.com/cs/)

Vixie P. and Schryver V. (1997) "*Distributed Checksum Clearinghouse"*. Referred on 26th September 2005 from World Wide Web: (www.rhyolite.com)

Voges, K. E. (2005). "*Research techniques derived from rough sets theory: Rough classification and rough clustering*". Paper presented at ECRM2005: Fourth European Conference on Research Methods in Business and Management, April 21 - 22, 2005.

Wang G.Y., Chen L. and Wu Y. (1999). "*Rough set based solutions for network Security*". IEEE Symposium on Security and privacy. IEEE Computer Society,1999.

Westbrook B. (2000). "*The Basic of Spam Filtering*". Referred on 17th October 2005 from World Wide Web: (http://www.mail-filters.com/)

Zhang L., Zhu J.B. and Yao T. (2004). "*An Evaluation of Statistical Spam Filtering Techniques*". ACM Transactions on Asian Language Information Processing (TALIP), v.3 n.4, p.243-269, December 2004.