SPAM BLOG PREVENTION WITH CONTENT ANALYSIS AND USER BEHAVIOUR MODEL

MOHAMMAD HAFIZ BIN ISMAIL

UNIVERSITI UTARA MALAYSIA

PREVENTING SPAM BLOGS USING CONTENT ANALYSIS AND USER BEHAVIOUR MODEL

A thesis submitted to Faculty of Information Technology in partial fulfillment of the requirements of the degree

Master of Science (Information Technology)

Universiti Utara Malaysia

By

Mohammad Hafiz bin Ismail

PUSAT PENGAJIAN SISWAZAH
(*Centre For Graduate Studies*)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

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

MOHAMMAD HAFIZ ISMAIL

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

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

PREVENTING SPAM BLOGS USING CONTENT ANALYSIS AND USER BEHAVIOR MODEL

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 field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **DR. FAUDZIAH AHMAD**

Tandatangan
(Signature) : _____

Tarikh
(Date) : 6/12/07

"to my beloved parents"

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the 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(s) 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.

Requests 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

Department of Computer Science

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Arnan.

ABSTRACT

Spam blog is a subset of blog which contains nothing more than stolen materials and inauthentic text designed to gain profit from various type of advertisements. Splogs have become a nuisance in the blogosphere because it pollutes search engine results and blog update servers. This paper discusses the similarity between spam blogs and email spams and the techniques used to identify them. The paper also propose the development of a prototype blog update server that implements content analysis and user behaviour model to filter splogs before they are indexed into blog search engine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATION

API	Application Programming Interface
OOAD	Object Oriented Design Methodology
PHP	PHP : Hypertext Preprosessor
RPC	Remote Procedure Call
RFC	Request For Comments
RSS	Really Simple Syndication
SURBL	Spam URL Realtime Black List
W3C	World Wide Web Consortium
XML	Extensible Markup Language

http://mypapit.net/

CHAPTER 1

INTRODUCTION

1.1 Background

Weblog or blog is a type of web site used as a publishing platform in the internet. Blog is distinguished from other type of web site as its content listed by date in reverse chronological order (Kolari et al., 2007). The term weblog was introduced by Jorn Barger in 1997 as "a Web page where a weblogger logs" (Blood, 2004), since then the term blog has entered popular usage as a noun and verb. A blog is typically used for publishing journal entries, online diary, news and personal notes in which it is updated regularly by its owner who is known as blogger. Blog then saw an increase of popularity, where the number of blogs has rose up to 5 million in 2003 compared to few hundreds in 1999. The term blogosphere were later introduced to refer to a community of blogs in a certain area of interests (Blood, 2004)

The contents of the thesis is for internal user only

REFERENCES

Bayes, T. (1763). *An Essay Towards Solving a Problem in the Doctrine of Chances*. Reprinted in: Bayesian Statistics: Principles,Models, and Applications.

Blood, R. (2004). *How blogging software reshapes the online community*. Communications of the ACM Volume 47 (12), 53-55

Fetterly, D., Manasse, M., Najork, M. (2004). *Spam, Damn Spam, and Statistics: Using statistical analysis to locate spam web pages*. ACM International Conference Proceeding Series; Vol. 67, Proceedings of the 7th International Workshop on the Web and Databases.

Fuchun, P., Dale, S., Shaojun, W. (2003). *Language and task independent text categorization with simple language model*. Proceedings, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, 189-196

Gili, K. E. (2005). *Blogging, RSS and the Information Landscape: A Look At Online News*. Workshop on the Weblogging Ecosystem.

Graham, P. J. (2002). A Plan For Spam. *Hackers and Painters* (pp. 109-117). Cambridge, MA: O'Reilly Media.

Gomes, L.H., and Castro, F. D. O., Almeida, V. A. F., Almeida, J. M, Almeida, R. B., Bettencourt, L.M.A. (2005). *Improving spam detection based on structural similarity*. Proceedings of the Steps to Reducing Unwanted Traffic on the Internet on Steps to Reducing Unwanted Traffic on the Internet Workshop, Cambridge, MA, p 12.

Graham, P. J. (2003). *Better Bayesian Filtering*. Retrieved March 28, 2007 from http://www.paulgraham.com/better.html

Graham-Cumming, J. (2006). POPFile Automatic Email Sorting using Naive Bayes. Retrieved September 28, 2007 from http://popfile.sourceforge.net/old.html

Hammersley,B. (2003). *Content Syndication with RSS*. O' Reilly & Associates, Inc.

Han, S., Ahn, Y. Moon, S. Jeong, H. (2006). Collaborative Blog *Spam Filtering Using Adaptive Percolation Search.*

Hassan-Montero, Y., & Herrero-Solana, V. (2006).Improving tag-clouds as visual information retrieval Interfaces.University of Granada, Faculty of Library and Information Science, Colegio.

Herkshop, S., & Stolfo, S. J. (2004). Identifying spam without peeking at the contents. *Crossroads: The ACM student magazine.*

Herkshop, S., Stolfo, S. J. (2005). *Combining email models for false positive reduction.* Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM Press.

Hovold, J. (2004). *Naive Bayes Spam Filtering Using Word-Position-Based Attributes*. Department of Computer Science, Lund University.

Kallen, I. (2006). *Method and apparatus for identifying and classifying network documents as spam.* United States Patent 20070078939.

Khan, O. (2006). *LDARank: Bringing Order to the Blogosphere.* CS294-10, Practical Machine Learning.

Kolari, P., Finin, T., Java, A. & Joshi, A. (2007). Towards Spam Detection at Ping Servers. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007).*

Kolari, P., Java, A., Finin, T., Mayfield, J., Joshi,A., & Martineau, J.(2006a). Blog track open task: Spam blog classification. *TREC 2006 Blog track notebook.*

Kolari, P., Java, A., Finin, T., Oates,T., & Joshi,A. (2006b). Detecting spam blogs: A machine learning approach. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*

Lin, Y. Sundaram, H., Chi, Y., Tatemura, J. Tseng, B. L. (2006). *Splog detection using self-similarity analysis on blog temporal dynamics*. ACM International Conference Proceeding Series; Vol. 215, p 1-8.

Macdonald, C., & Ounis, I. (2006). The TREC Blogs06 collection : Creating and analysing a blog test collection. Department of Computing Science University of Glasgow Scotland, UK.

Manavoglu, E., Pavlov, D., Giles, C. L. (2003). *Probabilistic User Behavior Models.* Proceedings of the Third IEEE International Conference on Data Mining, p 203.

McCallum,A., Nigam, K. (1998). *A comparison of event models for Naive bayes text classification.* AAAI-98 Workshop on Learning for Text Categorization.

Moor, A., Efimova, L. (2004). *An argumentation analysis of weblog conversation.* Proceedings of the 9th International Working Conference on the Language-Action Perspective on Communication Modelling.

Ntoulas, A., & Najork, M. (2006). Detecting Spam Web Pages through Content Analysis. *In Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, 83-92

Osmar, R. Z., Antonie, M. (2002). *Classifying text documents by associating terms with text categories.* Proceedings of the 13th Australasian database conference, Volume 5, p 215-222

Pang-Ning, T., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining.* Massachusetts, Boston : Pearson Education

Pollit, M. (2005, November 2005). Cashing in on fake blogs. *The Guardian.* Retrieved March 23, 2007 from http://technology.guardian.co.uk/weekly/story/0,16376,1643774,00.html

Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.(2003) *Tackling the poor assumptions of Naive Bayes text classifiers.* In Fawcett, T., Mishra, N., eds.: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, D.C., AAAI Press (2003) 616--623

Rish, I., Hellerstein, J., Jayram, T. (2001). *An analysis of data characteristics that affect Naive Bayes performance.* Proceedings of the Eighteenth Conference on Machine Learning.

Rullo, P., Cumbo, C., Policicchio, V. L. (2007). *Learning rules with negation for text categorization*. Symposium on Applied Computing, Proceedings of the 2007 ACM symposium on Applied computing, 409-416.

Satzinger, J. W., Jackson, R. B., & Burd, S. D. (2004). *System analysis and design in a changing world*. Massachusetts, Boston : Course Technology.

Salvetti, F., & Nicolov, N. (2006). *Weblog classification for fast splog filtering: A url language model segmentation approach*. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, 137–140.

Shen, Y., Jiang, J. (2003). *Improving the Performance of Naive Bayes for Text Classification*. CS224N Spring 2003.

Stern, H., Mason, J. & Shephard, M. (2004). *A linguistics-based attack on personalised statistical E-mail classifiers*. Faculty of Computer Science Dalhousie University.

SURBL. (2004). *Introduction : SURBL Spam URL Realtime Black List*. Retrieved September 26, 2007 from http://www.surbl.org/introduction.html

Stolfo,S.J, Wei-Jen Li,Hershkop, S., Wang, K., Nimesker, O. (2003). *Detecting Viral Propagations Using Email Behavior Profiles*. Columbia University.

The Internet Society (2005). *The Atom Syndication Format*. Retrieved September 26, 2007 from http://atompub.org/rfc4287.html#rfc.section.1

UMBC ebiquity. (2006). *Splog software from hell*. Retrieved March 23, 2007 from http://ebiquity.umbc.edu/blogger/splog-software-from-hell/

Wei, K. (2003). *A Naïve Bayes Spam Filter*. CS281A Project.

Winer, D. (2001). *Weblogs.Com XML-RPC interface*. Retrieved March 21, 2007 from http://www.xmlrpc.com/weblogsCom

Wikipedia. (2006). *Spam (Electronic)*. Retrieved March 21, 2007 from http://en.wikipedia.org/wiki/Spam_(electronic).

Yu-Ru, L., Wen-Yen, C., Xiaolin, S., Sia, R., Xiaodan, S., Yun, C., Koji, H.,

Sundaram, H., Tatemura,J., & Tsen, B. (2006). *The Splog detection task and a solution based on temporal and link properties*. NEC Laboratories America.