ISLAMIC WEB PAGES FILTERING AND CATEGORIZATION

NURFAZRINA BINTI MOHD ZAMRY

UNIVERSITI TEKNOLOGI MALAYSIA

ISLAMIC WEB PAGES FILTERING AND CATEGORIZATION

NURFAZRINA BINTI MOHD ZAMRY

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

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2013

“To my Beloved Big Happy Family and Wonderful Best- Friends”

ACKNOWLEDGMENT

Bismillahiramanirahim. First and foremost, praise to Allah s.w.t to give me strength and His blessing to complete this Project. Special thanks to my supportive supervisor, Dr. Anazida Zainal for her guidance, advice and knowledge. Without her encouragement and patience this research would not have been the same as presented here. Not to forget, thanks to all the lecturers for the advice.

Thank you very much Malaysia Government for sponsoring my Master study. A bunch of appreciation to the Postgraduate staff for their helps. A lot of loves to my bestfriends, who are always there for me when I'm down. Thanks to my fellow classmate for the support and information sharing.

Last but not least, a deepest love to my beloved family for the courage and endless prayers for me. May Allah bless you.

ABSTRACT

The Internet creates the world without boundaries where people can get lots of information just by surfing the Internet. But still some of the information is not genuine and correct. Because of that, some of the practitioners of deviant teachings can take this opportunity to attract followers just using the Internet especially to distort beliefs of Muslim in Malaysia. Web filtering can be used as protection against inappropriate and prevention of misuse of the network, hence, it can be used to filter the content of suspicious websites and alleviate the dissemination of such website. Currently, process for blocking the deviate teaching website is done manually and in addition there are limited web filtering product offered to filter religion content and very limited for Malay language. This project is aim to classify deviant teachings Website into three categories which is deviate, suspicious and clean. Pre-processing, feature selection and classification are process involved in Web filtering process. In pre-processing three processes are involved: HTML parsing, stemming and stopping to produce the deviant teaching keyword. Three existing term weighting scheme namely TF, TFIDF and Modified Entropy are used as feature selection process in filtering deviant teaching website while Support Vector Machine (SVM) will be used for classification process. Classification is validated by accuracy, precision, recall and F1. 300 Web pages were collected from Internet based on three categories: deviant teaching, suspicious and clean Web pages. As a result, M.Entropy shows the most suitable term weighting scheme to use in Islamic web pages filtering rather than TFIDF and Entropy.

ABSTRAK

“Internet mewujudkan dunia tanpa sempadan di mana orang ramai boleh mendapatkan banyak maklumat hanya dengan melayari Internet. Namun, beberapa infomasi dalam talian ini belum pasti yang asli dan benar.. Oleh itu, beberapa pengamal ajaran sesat mengambil peluang ini untuk menarik pengikut hanya menggunakan Internet terutamanya untuk memutarbelitkan kepercayaan Islam di Malaysia. Oleh kerana Penapisan Web digunakan untuk perlindungan terhadap laman Web yang tidak berpatutan dan pencegahan penyalahgunaan rangkaian, maka, sistem penapisan Web boleh digunakan untuk menapis kandungan laman-laman web yang mencurigakan dan mengurangkan penyebaran laman web itu. Pada masa ini, proses untuk menapis laman Web ajaran sesat dilakukan secara manual serta produk penapisan web begitu terhad untuk menapis kandungan agama dan sangat terhad dalam Bahasa Melayu. Projek ini bertujuan untuk mengelaskan laman Web ajaran sesat kepada tiga kategori iaitu sesat, mencurigakan dan bersih. Pra-pemprosesan, *feature selection* dan klasifikasi adalah proses yang terlibat dalam proses penapisan Web. Dalam pra-pemprosesan tiga proses terlibat: *HTML parsing, stemming* dan *stopping* bertujuan untuk menghasilkan kata kunci ajaran sesat. Tiga *term weighting scheme* yang sedia ada iaitu TFIDF, *Entropy* dan *Modified Entropy* digunakan sebagai *feature selection* dalam penapisan laman Web ajaran sesat manakala *Support Vector Machine* (SVM) digunakan untuk proses pengelasan. Proses pengelasan disahkan oleh ketepatan (*Accuracy*) , ketepatan (*Precision*), ingat (*Recall*) dan F1. 300 laman web telah dikumpulkan dari Internet berdasarkan tiga kategori: laman web ajaran sesat, mencurigakan dan bersih. Hasilnya, *M.Entropy* menunjukkan skim pemberat istilah yang paling sesuai untuk digunakan dalam laman web Islam penapisan berbanding TFIDF dan *Entropy*.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURES

LIST OF ABBREVIATION

ARPANET	Advanced Research Projects Agency Network
CPBF	Class Profile Based Feature
CUDA	Compute Unified Device Architecture
GA	Genetic Algorithm
GNU	GNU's Not Unix
HTML	Hyper Text Markup Language
IE	Internet Explorer
IR	Information Retrieval
ISP	Internet Service Provider
JAKIM	*"Jabatan Kemajuan Islam Malaysia*" (Department of Islamic Development Malaysia)
JAPAS	*"Jawatankuasa Menangani Ajaran Sesat Peringkat Kebangsaan"*
JAPPIS	*"Jawatankuasa Penyelaras Penyelidikan Islam Peringkat Kebangsaan"*
LISP	List Processing
MoHA	Ministry of Home Affair
PCP	Parallel Coordinate Plot
PHP	PHP: Hypertext Preprocessor
PICS	Platform for Internet Content Selection
PSO	Particle Swarm Optimization
SQL	Structured Query Language
SVM	Support Vector Machine
TFIDF	Term Frequency - Inverse Document Frequency
URL	Uniform Resource Locator
VSM	Vector Space Model
WEKA	Waikato Environment for Knowledge Analysis
WWW	World Wide Web

CHAPTER 1

INTRODUCTION

1.1 Introduction

Since the Internet was introduced, it brings a very high impact to people around the world. The Internet creates the world without boundaries where people can get much information just by surfing the Internet. Internet comes from the term Internet working, connecting computer network using specific gateway or router. In the beginning, the ideas of the Internet were from the research of development of ARPANET by United State department of Defence. Nowadays, Internet became faster compare to the last twenty years which offers thousands of information whether it was true or false. Through the Internet, people can interact among each other by search for information, sending e-mail, perform electronic transaction and even social networking.

One of the biggest challenge people face from the Internet is the accuracy of content in the Internet. Not all of this online content is accurate, pleasant, or inoffensive (Heins *et al*., 2006). As the Internet can be accessed from people around the world in faster manner, some of the practitioners of deviant teachings can take this opportunity to attract followers just using the Internet especially to distort beliefs of Muslim in Malaysia. There are some terminologies which this group use to disseminate their doctrine where it is obviously deviate from the Islamic teaching. Some Muslim which does not have a solid foundation of Islam may simply be attracted to this group. Web filtering has provided two major services: protection

against inappropriate and prevention of misuse of the network (Du *et al*., 2002). Hence, the web filtering system can be used to filter the content of suspicious websites and alleviate the dissemination of such website. Moreover some of the content may threaten the society as there will reflecting in violence, hate as well as undermine national harmony.

1.2 Problem Background

Web filtering is commonly used by organizations such as offices and schools to prevent computer users from viewing inappropriate web sites or content, or as a pre-emptive security measure to prevent access of known malware hosts. While web filtering software, a term for software designed and optimized for controlling what content is permitted to a reader, especially when it is used to restrict material delivered over the Web. Early filtering was based on either "self-rating" by online publishers or "third-party rating" by filter manufacturers (Heins *et al*., 2006).

Filtering can be divided into five types which include Browser based filters, Client-side filters, Content-limited ISPs, Network-based filtering and Search-engine filters. Each of this filtering has their own characteristic which can be used by different level of users. Browser based filters usually are implemented by third party browser extension while Content-limited ISPs mostly used by government regulator and parental control as it have offered by ISP went subscribe their services. In Client-site filters, specific software are installed in the client computer which administrator can managed the software itself. Network-based filtering is enforced at the transport layer or application layer where the entire user within the network is bound to the organization policy. Lastly, Search-engine filters may be installed in the search engine like Google and Yahoo where it can be activated to block the certain website when safety filter are activated.

There are four types of web filtering approaches mentioned by Lee *et al*., (2006). There is Platform for Internet Content Selection (PICS), Uniform Resource Locator (URL) blocking, keyword filtering and intelligent content analysis. From those four approaches intelligent content analysis is the best to classify the web content since it can be used to categorize Web pages into different groups for instance the pornographic and non-pornographic. Moreover keyword filtering sometimes will over-block the Web while URL blocking keeps maintaining the reference list.

The self-regulating nature of the Web community, coupled with the easy making information available on the Web has led some individuals to abuse their freedom of expression by putting up harmful materials on the Web (Lee *et al.,* 2003) and this includes practitioners of deviant teachings to disseminate their doctrine. In Malaysia, Malaysian Department of Islamic Development (JAKIM) is responsible to monitor all Islamic oriented publications in various fields where their objective are to develop the Islamic community mind and spreading Islamic *da'wah* through Islamic oriented printed publications in various fields.

Currently, process for blocking the deviate teaching website is done manually. The processes start when JAKIM receive a report from people either by telephone or via email. Special unit in JAKIM then investigate the reported website to filter the website whether it may contain the deviant teaching issues where the investigation normally last for some months. Since the process will take more time, hence by the mean time some of Internet user may accept doctrine. One of the most popular groups of deviant teachings is spiritual and mystical theosophy or in Malay it was called *"Batiniah dan Tasawuf Teosofi"*.

There are many classification method used to classify the Website into respective group. For instance, Naïve Bayes, Decision Tree, Artificial Neurel Network as well as Support Vector Machine are methods used for Website classification. In this project, Support Vector Machine also known as SVM is chosen as classification method to classify the Website into specific group. SVM is the

supervised classification method composed of training data and testing data which used for learning and classifying process respectively. Moreover, SVM is suitable to use in text classification and more effective than other methods.

1.3 Problem Statement

The current passive filtering approach used to block the deviant teachings Website impact on the effectiveness for blocking the website. However, in Malaysia there is no web filtering product offered to filter this kind of religion content especially in Malay language. Furthermore, blocking the Web sometime may cause of misunderstanding when there are some of the Web are intent to educate the readers. The effect of this uncontrolled disseminate of deviant teachings through the website may also result of distort beliefs to Muslim society especially in Malaysia.

1.4 Project Aim

The aim of this project is to classify the deviant teachings Website into three categories which is deviate, suspicious and clean. Nevertheless, by implementing this Malay language Web filtering system may help specific organizations to control the disseminated of deviant teachings over the Internet. In the meantime this project can help to minimize spreading of the deviant teachings through the website and hence improve the distort beliefs to Muslim society.

1.5 Project Objective

The main objectives of the project have been identified in order to complete this research, there are;

i. To pre-process and identify the deviant teaching keywords from the Web pages.
ii. To implement the features selection method for deviant teaching terms and classifies the term using SVM.
iii. To test and validate the accuracy of the classification of deviant teaching.

1.6 Project Scope

In accomplishing this project, a number of scopes have been determined, which are:

i. The samples of Web pages will be extracted from the Internet.
ii. Classification will be implemented into three categories which include of deviate, suspicious and clean.
iii. The Support Vector Machine (SVM) algorithm will be used to classify Web content.
iv. The Web pages used during the project is only in Malay language.

1.7 Significance of Project

This project is prepared to accomplish the objectives to filter and categorize the Islamic Web specific on deviant teachings pages by implementing the Support Vector Machine algorithm. Thus, this project is important to categorize the level of

deviant teachings to aware the readers especially Muslim since the content may lead to distorted of their beliefs. Hence it can minimize the spread of the deviant teaching Furthermore the study can help organization like JAKIM, school, university community to filter their web content within the networks.

1.8 Project Organization

This chapter has covered about the introduction of the project including the problem background, problem statement, project aims, project objectives, project scopes, and significance of the project. This project will be continue with Chapter 2 that discuss more on literature review of the project collected from some resources such as Internet, books, journals and collecting primary data using different research methods. Chapter 3 will covered the project methodology where it highlights the methodology used in this study and discuss about the data used during this study. Chapter 4, the implementation of pre-processing which highlight the second objective is discussed. Next, Term Weighting Schemes in feature selection process and classification of Web pages in SVM will be discussed in detail. Lastly, chapter 6 will concludes the overall of this project.

REFERENCES

Ahmad, F., Yusoff, M. and Sembok T. M. T. 1996. *Experiments with a Stemming Algorithm for Malay Words*, Journal of the American Society for Informatlon Science. 4712. pp. 909-918.

Barry, M. L., Vinton, G. C. and David, D. C. 1997. *The Past and Future History of the Internet*, Communications Of The ACM, Vol. 40, No. 2, pp. 102-108.

Barry, M. L., Vinton, G. C., David, D. C. 2009, *A Brief History of the Internet*, ACM SIGCOMM Computer Communication Review. Vol.39, pp.22-31.

Blum, A. and Langley, P.1997. *Selection of Relevant Features and Examples in Machine Learning*. Artificial Intelligence. 971-2. pp. 245–271.

Chang, C. and Lin, C.J. 2012. *LIBSVM: A Library for Support Vector Machines*. Department of Computer Science National Taiwan University, Taipei, Taiwan

Chang, C. and Lin, C.J. 2012. *LIBSVM -- A Library for Support Vector Machines Website* [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [Accessed December 2012].

Chen, R. C. and Hsieh, C. H. 2006. *Web page classification based on a support vector machine using a weighted vote schema*, Systems with Applications. Vol. 31. pp. 427–435

Cummins, R., and O'Riordan, R. 2007. *Evolved Term-Weighting Schemes in Information Retrieval: An Analysis of the Solution Space*, Dept. of Information Technology, National University of Ireland, Galway,Ireland

Du, R., Safavi-Naini, R. and Susilon W. 2003. Web filtering using text classification, The 11th IEEE International Conference on Networks. pp.325-330.

Gupta, S., Kaiser, G., Neistadt, D. and Grimm P. 2003. *DOM-based content extraction of HTML document*, In Proceedings of the 12nd international conference in World Wide Web, ACM New York, NY, USA, 207-214

Guyon, I. and Elisseeff, A. 2003. *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research. Vol. 3. pp.1157-1182

Finkelstine, S. 2002. *BESS's Secret LOOPHOLE censorware vs.privacy & anonymity*.[Online].Available:http://sethf.com/anticensorware/bess/loophole.php. [Accessed 18/10/12].

Heins, M., Cho, C. and Feldman, A. 2006. *Internet Filters, A Public Policy Report*. Brennan Center For Justice, NYU School Of Law, New York, United State of America.

Ismail, S.Z. 2010, *"Menangani Ajaran Sesat Di Kalangan Umat Islam: Perspektif Undang-Undang Dan Pentadbiran"*. Shariah Journal, Vol. 18, No. 2 (2010) pp. 247-276.

Kohavi, R. and John G.H. 1997. *Wrappers for feature subset selection*. Artificial Intelligence 97. pp. 273-324.

Lee, P.Y., Hui, S.C. and Fong A.C.M 2002. *Neural Networks for Web Content Filtering*, IEEE Intelligent Systems. pp. 48 - 57

Lee, P.Y., Hui, S.C., Fong, A.C.M. 2003. *A structural and content-based analysis for Web filtering*, Internet Research. Vol. 13 Issue: 1 pp. 27 - 37

Lee, Z. S., Maarof, M. A., Selamat, A., and Shamsuddin S. M. 2008. *Enhance TermWeighting Algorithm as Feature Selection Technique for Illicit Web Content Classification*, Eighth International Conference on Intelligent Systems Design and Applications ISDA. Nov. 26-28, Kaohsiang City, Taiwan, pp. 145-150

Lee, Z. S. 2010. *Enhanced Feature Selection Method For Illicit Web Content Filtering*. Ph.D Thesis, Universiti Teknologi Malaysia

Liu, Y., Loh, H. and Sun, A. 2009. *Imbalanced text classification: A term weighting approach.* Expert Systems with Applications. 361, 670-701.

Lin, P. C., Liu, M.D., Lin, Y.D., And Lai, Y. C. 2008. *AcceleratingWeb Content Filtering by the Early Decision Algorithm*. IEICE TRANS. INF. & SYST. Vol.E91–D, No.2, pp-251-257.

Mazlam, N.F. 2012. *Enhancement of Stemming Process For Malay Illicit Web Content*. Master Thesis, Universiti Teknologi Malaysia.

Othman, A. 1993. *Pengantar perkataan Melayu untuk sistem capaian dokumen Introduction to Melayu Words for Document Retrieval System*. M.S. thesis, University Kebangsaan Malaysia.

Lee, P.Y., Lee, S.C. and Fong, A.C.M. 2003. *A Structural And Content-Based Analysis For Web Filtering*, Internet Research. Vol. 13 Issue: 1, pp. 27 – 37, Sebastian, F. 2002. Machine learning in automated text categorization. ACM Computing Surveys, Vol 34, 1–47.

Raghavan, V. and Wong, S. 1986. *A critical analysis of Vector Space Model for Information Retrieval*. Journal of the American Society for Information Science. 35(5)

Salleh, S.F. 2012. *Comparative Study On Term Weighting Schemes As Feature Selection Method For Malay Illicit Web Content Filtering Content*. Master Thesis, Universiti Teknologi Malaysia.

Salton, G., Wong, A. and Yang, G. 1975. *A Vector Space Model for automatic Indexing*

Selamat, A. and Omatu, S. 2004. *Web Page Feature Selection and Classification Using Neural Networks*, Information Sciences. 158 , pp. 69–88

Sembok, Tengku M.T., Uman, I.I.H. and Rocessing, I.N. 2005. *Word Stemming Algorithms and Retrieval Effectiveness in Malay and Arabic Documents Retrieval Systems*, Conference on Research and Development in Information Retrieval. pp.95–97.

Verikas, A. and Bacauskiene M. 2002. *Feature Selection With Neural Networks*, Pattern Recognition Letters , 23,1323–1335

Wikipedia*, "Contentcontrolsoftware*",http://en.wikipedia.org/wiki/Contentcontrol _software visited 7/8/12.

Wikipedia, "*Web Crawler*" http://en..org/wiki/Web_crawler visited 18/10/12.

Wikipedia, "*tf–idf*" http:// http://en..org/wiki/Tf%E2%80%93idf visited 18/10/12.

Yang, Y. and Liu, X. 1999. *A Re-Examination Of Text Categorization Methods*, Proceedings Of The 22nd International ACM SIGIR Conference On Research And Development In Information Retrieval. pp. 42–49.