ENHANCEMENT OF STEMMING PROCESS FOR MALAY ILLICIT WEB CONTENT

NOOR FATIHAH BINTI MAZLAM

UNIVERSITI TEKNOLOGI MALAYSIA

ENHANCEMENT OF STEMMING PROCESS FOR MALAY ILLICIT WEB CONTENT

NOOR FATIHAH BINTI MAZLAM

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 Computer Science and Information Systems
Universiti Teknologi Malaysia

AUGUST 2012

Dedicated to my beloved parents; Abah and Mama, and my brothers; Muhammad Alif and Muhammad Khalaf.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful

First and foremost, all praises to Allah for all the strengths and His blessings in completing this thesis. Special appreciation goes to my supervisor, Dr Anazida bt Zainal, for her supervision and constant support. This thesis would not have been possible without the guidance and advice from her.

Besides, I would like to thank the Malaysia Government and Universiti Malaysia Pahang for sponsoring my Masters study in Universiti Teknologi Malaysia. I also would like to express my appreciation to all my classmates and my bestfriend, Siti Nur Shahida Che Kar for their support and help during my study.

Last but not least, my deepest gratitude goes to my beloved parents and my brothers, for their endless prayers, love and encouragement. May Allah always shower His blessings to our family.

ABSTRACT

Web filtering system is one of the systems use to prevent users from can access any web pages that contain illicit contents. There are six (6) phases included in web filtering process. One of them is pre-processing phase. In this phase, there are three main activities included; HTML parsing, stemming, and stopping. The main focus in this research is stemming process. Stemming process is used to remove any affixes that attached together in the input words from web pages to produce the correct root words. To date, the existing stemming algorithm in Malay language; Othman's stemming algorithm and Sembok's stemming algorithm still produce errors in the result. Hence, the errors from both stemming algorithm were analyzed. Few features were created to encounter the problems occurred in existing stemming algorithm. There are initial checking with dictionary, implementation of Rule 2 and also checking with additional dictionary that contains the illicit words not included in the initial dictionary. These new features were added in enhanced stemming algorithm.In order to check the effectiveness of the new features added in the enhanced stemming algorithm, few tests were done to the sample of web pages. Based from the test, the result shows that only 11% corrected words produced if the test is done by without checking with initial dictionary and 72% corrected words produced if the process starts with initial checking with dictionary. The result for the test for implementation of Rule 2 shows that by using Sembok's algorithm it produced only 17% corrected words compared with enhanced stemming algorithm produced 62% corrected words. As conclusion, the implementation of new features in enhanced stemming algorithm can reduce the errors produce in Sembok's stemming algorithm.

ABSTRAK

Sistem web penapisan adalah merupakan salah satu system yang digunakan untuk mengelakkan pelayar Internet daripada melayari mana mana laman web yang mengandungi bahan yang tidak bersesuaian. Di dalam system web penapisan, ianya terdapat enam fasa. Salah satu daripadanya adalah fasa pra-pemprosesan. Di dalam fasa ini, terdapat tiga aktiviti utama iaitu penghuraian kod HTML, pembuangan imbuhan pada perkataan, dan juga pembuangan perkataan yang tidak berkait semasa proses pencari perkataan di dalam laman web. Fokus utama di dalam kajian ini adalah merupakan pada aktiviti pembuangan imbuhan yang terdapat pada sesebuah perkataan untuk menghasilkan kata akar yang betul. Setakat ini, hanya terdapat dua algoritma yang ada digunakan untuk perkataan di dalam Bahasa Melayu, iaitu Algoritma Othman dan juga Algoritma Sembok. Kesalahan yang terdapat di dalam kedua algoritma ini telah dikaji dan beberapa cadangan untuk penambahbaikan untuk algoritma telah dilakukan. Antara cadangan atau langkah yang telah diambil adalah dengan memeriksa input perkataan terlebih dahulu dengan kamus di dalam sistem, pelaksanaan Rule 2, dan juga menambah satu kamus yang mengandungi perkataan yang tidak terdapat di dalam kamus terdahulu. Berdasarkan eksperimen yang dibuat, ia menunjukkan hanya sebanyak 11% kata akar yang betul dihasilkan jika pemeriksaan input perkataan dengan kamus dijalankan berbanding 72% kata akar yang berjaya dihasilkan jika input perkataan disemak dengan kamus terlebih dahulu. Untuk perlaksanaan Rule 2, hanya 17% kata akar yang betul sahaja berjaya dihasilkan dengan menggunakan Algoritma Sembok berbanding 62% kata akar yang betul berjaya dihasilkan dengan menggunakan algoritma yang baru. Kesimpulannya,perlaksanan cadangan dan kaedah untuk algoritma yang baru ini berjaya untuk mengurangkan kesalahan yang terdapat di dalam Algoritma Sembok.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

LIST OF ABBREVIATION

CPBF	Class Profile Based Feature
HTML	HyperText Markup Language
M.Entropy	Modified Entropy
SVM	Support Vector Machine
TF	Term Frequency
URL	Uniform Resource Locator
VSM	Vector Space Model

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

1.1 Introduction

In recent years, Internet has become widely used for many purposes to the individuals. Most of the common purposes the using of the Internet is to look for the information, email, online shopping and also social networking. But, not all the contents from the Internet are useful to the society, especially for the children. Illicit web content such as pornography, bullying, violence and so on can give such a bad impact to their mental health and also might bring them to involve in any extremely dangerous violent desire. Easy access to this harmful content of web pages is one of the factor parents need to monitor their children Internet surfing activities. A proper system that can help to block these unhealthy web pages is needed to tackle this problem. It also can help parents to monitor their children Internet surfing activities.

One of the methods that can be used to block these illicit web pages is by using web-filtering system. Web-filtering system is a program that can screen the web page to determine whether some or all it should not be displayed to the user. Then, the filter checks the origin of the contents of the web pages according to the set of rules provided by the user to block the web pages that installed the Web filter. In this study, by using the web filtering system, it will block any web pages that contained pornographic content.

1.2 Problem Background

Web filtering content can be implemented in a few ways, either by installing a software program on the computer or by servers that providing the internet access. Besides that, Internet service provider (ISP) also offered the service for web content filtering. It blocks any illegitimate content in the web pages before it enters the network at home.

There are four types of web content filtering. There are client-side filters, content-limited ISPS, server-filters side, or search-engine filters. The usage of these filters depends on the organizations or the situation to be applied. For instance, client-side filters may consider to be installed at home because this filter can be customized to meet the family's need. It can be controlled or disabled only by one person who had the password for this software. For content limited ISPS customer only can access the set portion of Internet content that provide by the service provider. For search-engine filters, when the safety filter is activated, it will filter the links from the search engine.

During the process of web filtering content, it involves a few steps of procedures to analyse the web page. It starts with web data collection until the last phase which is to classify the pages either it contains the objectionable content or not. One of the phases during this technique is pre-processing. During this phase, it will extract the pages where only the text and images in page are included. So, for the text content, it will undergo the process for stopping and stemming. Stemming process is a process where it will removes the affixes that attached together on words to produce only the root word, while stopping is the process to eliminate the regular words in the document based on stop-list.

The problem when using these types of web content filtering system, not all this system can interpret the web pages that have objectionable content due to its weaknesses to match the input text with the list of keywords for pornographic term in the dictionary. This is where stemming process plays its role. The efficiency of the system filtering is depending on how the stemming algorithm can work effectively to remove the affixes to produce the root word. By doing so, system can detect any text

input that will lead the user to the any web pages that contains illicit or illegal contents.

But most of the stemming algorithms are built to suits with English words, which is not applicable to use in this study. It is because the samples for web pages in this study only using Malay language. The problem may arise because the structure of the words in Malay language. In English words, it can be very simple by only removing plurals, past, and present particles compared than Malay words. In Malay words, there are four class of affixes; consists of prefixes, suffixes, infixes, and the most frequent occur among Malay words is prefixes-suffixes pair.

1.3 Problem Statement

The main aim in this study is to find "*What is the suitable stemming algorithm used to truncate the word into the root word that will reduce the vocabulary size and improve recall*". Hence, below are the questions that related to the main question in this study.

i. Do the existing stemming algorithms can be used to conflate the morphological variants in Malay Language?
ii. Which implementation orders of the rules can reduce the error during stemming process?
iii. How to cater the keywords for pornography which are not included in the dictionary?

1.4 Project Objective

The objectives of this study are defined as below:

i. To enhance Sembok's stemming algorithm that suits for Malay words.
ii. To design the enhanced stemming algorithm that suits with the scope of this study
iii. To test and validate the enhanced stemming algorithm with Sembok's stemming algorithm.

1.5 Project Scope

The scopes which will identify the boundary for this study are:

i. The study only focuses on the stemming algorithm technique in the process of web content filtering.
ii. The samples for web pages used in this study are obtained from the Internet.
iii. The language used in the web pages is only using Malay language.

1.6 Motivation and Significant of Project

As the information on the Internet are much easier to access to anyone especially the children, parents should aware the contents of the web pages surfed by their children. The inappropriate contents from web pages such as web pornography can give bad effect to them. Thus, web content filtering method should be used to filter or distinguish the content of the web pages either it is useful or risky content. So, this software is useful to use to block any risky web pages.

Besides, this software also has potential to be implemented at the school environment. Teachers or school administrator also have their own responsibility to make sure facilities for internet access at the lab computers are using precisely by the

students. So, in order to make sure the students access the right web pages, this web content filtering is a right decision to install it in the lab computers.

1.7 Project Organization

This study will covers six chapters. Chapter One describes on the problem background of the study, project objectives, project scopes, the significant of the study and also the chapter organisations for this study.

Chapter Two of this study will explain on the literature review. It will review the web filtering systems currently used and also its evolutionary and also the technology trend for filtering approach. Then, it will discuss about the stemming algorithm that applied in the web filtering content. It also will focus on the current research on this stemming algorithm, what is missing in this research, and what need to include to further research in this area.

Chapter Three will explain on the research methodology for this study. It includes the entire project framework that will describe all the phase in this study. Chapter Four will discuss about analysis conducted in early phase of this study, including the comparisons of existing stemming algorithm mentioned in the literature review, Chapter Two.

Then, the results and analysis of the comparisons between Sembok's stemming algorithm and enhanced stemming algorithm will be discussed in the Chapter Five. The conclusion of the project is included in the Chapter Six. The constraint of the project and recommendation for future works also include in Chapter Six.

REFERENCES

Ahmad, F., Yusoff, M. and Sembok, Tengku M T. (1996). Experiments with a Stemming -Algorithm for Malay Words, *Journal of the American Society for Information Science and Technology*. pp 909–918.

Al-shammari, E.T., (2008). Towards an error-free stemming.,*LADIS European Conference on Data Mining*. pp 160–163.

Asian, J., Williams, H. E., and Tahaghoghi, S. (2004). A testbed for Indonesian text retrieval. *In Proceedings of the 9th Australasian Document Computing Symposium (ADCS'04)*. University of Melbourne, Department of Computer Science, Melbourne, Australia. pp 55--58.

Dewan Bahasa Dan Pustaka. (2004). *Kamus Dewan* (Council Dictionary). Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.

Erk, K. & Pad, S., A Structured Vector Space Model for Word Meaning in Context. , *Proceedings of the Conference on Empirical Methods in Natural Language Processing* . pp 20-24.

Fox, C., (1999). A stop list for general text. *ACM SIGIR Forum*, pp.19–21.

Harman, D. (1991) How Effective is Suffixing. *Journal of the American Society for Information Science* 42 (1), 7—15

Hindrajaja L.S (2003) Automatic Learning on Stemming Rules for Indonesian Language. *Pacific Asia Conference on Language* . pp 325-328

Kadri, Y. & Nie, J., (1999). Effective Stemming for Arabic Information Retrieval, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. pp.68–74.

Kwee, A.T., Tsai, F.S. & Tang, W., (2009). Sentence-Level Novelty Detection in English and Malay.*Advanced in Knowledge Discovery and Data Mining* , pp.40–51.

Lee Z.S, (2010). Enhanced Feature Selection for Illicit Web Content Filtering.PhD thesis. Universiti Teknologi Malaysia (UTM)

Lin, S.-S., (2009). A document classification and retrieval system for R&D in semiconductor industry – A hybrid approach. *Expert Systems with Applications,*36(3). Pp 4753–4764.

Lovins, B., (1968). Development of a Stemming Algorithm. *Journal* , pp.22–31.

Molina, L.C. et al. (2001), Feature Selection Algorithms : A Survey and Experimental Evaluation*, .IEEE International Conference on Data Mining*. pp 306-313

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

Porter, M. F.(2003). An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems,* pp 130--137.

Robert Krovetz, (1993) Viewing Morphology as An Inference Process, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp 191-202

Savoy, J. (1993). Stemming of French words based on grammatical categories. *J. Amer. Soc. Inform.* Sci. 44, pp 1-9.

Sembok, Tengku Mohd T & Bakar, Z.A., (2011). Effectiveness of Stemming and n-grams String Similarity Matching on Malay Documents. *International Journal of Applied Science and Information Technology* .pp.208–215.

Sembok, Tengku Mohd T, Uman, I.I.H. & 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.

Solka, J.L., (2008). Text Data Mining: Theory and Methods. *Statistics Surveys,* 2, pp.94–112. Available at: http://projecteuclid.org/euclid.ssu/1216238228 [Accessed March 9, 2012].

W. Kraaij & R. Pohlman (1995). Evaluation of a Dutch Stemming Algorithm. *The New Review of Document and Text Management*, volume 1, pages 25–43.

Williams, H.E. & Tahaghoghi, S.M.M., (2005). Stemming Indonesian. *Proceedings of the Twenty-eighth Australasian conference on Computer Science* , pp 317-324.