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

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.

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.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, it starts with the general information regarding the issues about the usage of Internet and also Web filtering system. Then, it will continue to discuss about the stemming and stopping which is one of the crucial elements needed during filtering the illicit Web pages.

2.2 Internet

Internet is become one of the popular services in order to people to get information. It becomes a useful resource to get not only information, but also to use the online services, such as online banking, e-learning, online shopping and much more. The rapidity of the changing technology sometimes also can expose the user to any illicit content in the Internet. Not all the good information can be find in the cyber space. The risks of user to browse any illegal contents on the web are very high due to easy access to those web pages. The examples for illegal contents in the web pages are violence, pornography, or even the pages that contain the sensitive issues to the religion.

This can contribute a bad effect to human especially for the young society. Let's take the pornographic issue in the web pages for an example. This issue has

one of the highest ranking in the search request on the Web search engine. This issue should be taken as serious issues since it will contribute social problem in our society. As consequences, people need to have their own responsibility to overcome this problem as a society. One of the countermeasure should be taken is by using Web content filtering.

2.3 Web Filtering System

Web filtering systems is one of the solutions that can be used to filter any objectionable content in the Web pages. It will controls the content is allowed to the user. These types of filters can be installed in many different ways and very applicable to used it in the school environment, offices, or by parents to monitor their kids to access any illegal contents.

There are a lot of web filtering systems in the market nowadays and one of them is web content analysis. It uses the linguistic analysis, machine learning including image processing components. One of the procedures included during the analysis is by stemming and stopping process. It is one of the crucial parts of this process in order to get the effective result of the filtered web pages. A few products for web filtering systems shown in the table below

Table 2.1: Products Available for Web Filtering Systems

Products	Descriptions
NetNanny	- Compatible with Microsoft Windows and Mac OS and certain smartphones. - Restrict file sharing, block and filter websites and social networking and restrict usage to present limits.

	- Features for owners/admins such as; activity reports, remote management, monitoring of instant messages. - It can also keep passwords entered on specific user profiles of any kind on file and accessible to the administrator.
SurfWatch	- The first software that allowed users to block explicit content on the Internet. - Disallowed computers from accessing specified sites and by screening for newsgroups likely to contain sexually explicit material, SurfWatch was able to aid parents, educators and employers in preventing access to offensive material from a specific computer.
Secure Web SmartFilter	- As an Internet filter, blocking minors using the public computers from accessing inappropriate web content. - It will install on the server between the users and the open Internet so it makes it harder to bypass, though it is not uncommon for students with more extensive computer knowledge to attempt to bypass the system.

2.4 Stemming and Stopping Algorithm

Stemming process has played an important role in order to improve the correctness of information retrieval. During the process, it removes common words into the root words. The efficiency of information retrieval system is depending on the system comprehend the meaning of the word in the document because normally in the system, text document is controlled by indexing process (N.Idris *et al,* 2001). So, each of the root words or terms is tagged for these documents as an indicator to referring to the documents.

Stopping is a process to removes the common or regular words in the documents by using a stop-word list. Stop words are the words that unrelated during the searching purposes because they are frequently occurred such as "are". "is", "the", "or" and etc. The purpose of this process is to save the space and time during the searching process. Stop words will slump during the indexing time and ignored at search time. In addition, it helps to remove noise data too.

2.4.1 Stemming Algorithm

Stemming algorithm is a procedure to extract the words into the root words. It is done by removing the affix. Affix removal stemmers will extract the prefixes or suffixes of the words based on the transformation rules. Example for the words that have affixed shown in the table below.

Table 2.2: Example for Affixes Word

Root words	Affix
Eat	Eat+s = Eats
Eat	Eat+ing = Eating
Eat	Eat+en= Eaten

Hence, by cut off the affix, it will make query match process much easy to find the related stem words and also to perform the information retrieval task. Stemming process can help information retrieval task to perform efficiently due to the reducing of size of the data in the system. Morphological variance is also one of the reasons why stemming process is needed in the information retrieval system.

For instance, an English word “absorb” also has similar meanings with “absorbs”, “absorbing”, and also “absorbed”. All of these words have the same root “absorb” and these words are called conflations. So, by decreasing the morphological variants among the words can help the system during the query-match process.

2.4.2 Stemming Algorithm in English

Stemming process also has widely used in English language. In English, it is easy to conflate the words because usually the morphological variants has take place at the right end of the word. This approach makes it is easy separate the root word from its words but it will causes two majors types of errors during the stemming process. Over truncation may occurs when too short a stem remains after truncation and it will resulted in unrelated words being conflated to the same root. Secondly, under truncation, if any too short string is removed, it may result in related words being described by different things.

2.4.2.1 Lovins Algorithm

Lovins stemming algorithm was presented by Julie Beth Lovins in 1968. This algorithm will removes the endings of the word based on the longest match principle and common suffixes such as *SES, *ING and also *ATION. The removal of prefixes, where the strings has been added at the left hand end of word of root has not studied for English words compared than in Malay words because of their morphological variant.

In Lovins algorithm, it spurred the development of many subsequent algorithms and more generally the use of stemming as a general tool used in the information retrieval.The issues has arise in this algorithm when a word is presented for stemming in a dictionary based stemming algorithm, the right had end of the word is checked for the presence of any suffixes found in the list. If it's found to be present, it is removed, subject to a range of context sensitive rules that forbid. For instance, the removal of *ABLE from the word TABLE or from the word GAS it will removes *S. A range of recoding rules may be provided to enable the conflation of variants such as for the word FORGETTING and FORGET or between the word ABSORB and ABSORPTION.

2.4.2.2 Porter Algorithm

The Porter stemming algorithm is a conflation stemmer developed by Martin Porter in 1980. It is based on the idea that the suffixes in the English language are mostly made up of a combination of smaller and simpler suffixes. The Porter algorithm differs from Lovins type stemmers in two major ways. The first difference is a significant reduction in the complexity of the rules associated with suffix removal. The need for simplicity is exemplified by previous stemming algorithm, Lovins stemming algorithm.

In Lovins algorithm, it contains only 294 suffixes each of which is associated with one of 29 context sensitive rules that determine when that suffix can be removed from the end of a word, and also contains 35 recoding rules. Despite the large number of suffixes, relatively few of them are plural nouns and both the suffixes and the recoding rules suggest that the Lovins algorithm has been designed principally for the processing of scientific texts. The second difference is the use of a single, unified approach to the handling of context. Many of Lovins' context sensitive rules relate to the length of the stem remaining after the removal of a suffix: the minimal acceptable length is normally just two characters, with a consequent risk of significant over stemming.

2.4.3 Stemming Algorithm in Malay

As in other language, Malay Language also needs to have a comprehensive stemming algorithm that used for indexing and also in the purpose for retrieval of Malay text documents. There are two existing stemming algorithm used for Malay language, Othman's algorithm and Sembok's algorithm. However, these two existing stemming algorithm still produce a lot of error during the stemming process.

Stemming algorithm for English is cannot be implemented in the Malay text document retrieval because of the usage of affixes in the words (Sembok T, 2005). In Malay language, it has four class of affix class; prefix, suffix, prefix-suffix, and infix. Thus, this make Malay language is much more complex compared than English and other language. For English stemmer, it has been discover that the stemmers only removed for the suffix part in the words. Nevertheless, in Malay language, stemmer needs to remove for both prefix and suffix parts to produce the root words.

Table 2.3: Example For Affix Classes

Affix class	Affixes
Prefix	Ke Di Men Ber Ter Meng Pen Per
Suffix	i an nya kan kah lah
Prefix-Suffix	Ber - an Ber - kan Di - i Di - kan Men - kan Men - i Memper - i Se - nya
Infix	el em er in

Table 2.3 shows the examples of affixes in Malay Morphology. Affix is the verbal elements that connected to the word, either at the beginning or at the end of the word. If the element is attached at the beginning of the word, it is called as prefix, while if it is attached at the end of the word, it called as suffix. Infix is when the element is attached at the middle of the word. Besides that, one word can be attached by both prefix and suffix at the same time, called as prefix-suffix pair (N.Idris *et al*, 2001).

In Malay Morphology, for prefix class may have some problems with spelling variations. It is occurred when the first letter of the root word is removed through the addition of prefixes to the root words beginning with some letter. As a consequence, errors may happen during the stemming process. For instance, to insert a prefix “mem” to a root word “pukul”, the first letter of the root word “p” will drop in order to perform the word “memukul”. This case may happen to the other prefixes “pem” where it will remove “f” or “p” of the root words, “meng” or “peng” where the letter “k” will be eliminated, “meny” or “peny” where it will remove the letter of “s”, and also for “men” or “pen” stemmer will removes the letter of “t”.

2.4.3.1 Prefixes

A prefix is the verbal element that attached at the in front of word (+ prefix). There are a lot of common prefixes in the Malay language such as di, ke, se, beR, beL, meN, teR, peN, and including peR. But, it has an issue when the prefix is removing from the words that started with beR, meN, teR, peN and peR. It is because it will change the spelling of the words. For example, for the Malay word “*penyanyi*”, when the prefix “pen” is remove from the root word “*nyanyi*”. It is depends on the first letter of the words that appended it together. However, for the prefixes “*ke*”, “*se*”, and “*di*” do not change the structure when they combine with the root word. Table 2.4 below shows the example of prefixes attached with root word.

Table 2.4: Example for Prefix+

Prefix + Root word	**Words**
Di + makan	*Dimakan*
*Ke + lua*r	*Keluar*
*beR + lar*i	*Berlari*
meN + sara	*Menyara*
PeN + asas	*Pengasas*

2.4.3.2 Suffixes

A suffix is the verbal elements joined together at the end of the root word. It has dissimilarity from prefixes which it do not have any changes in word spelling when suffix is remove from the root word. Example of the common suffix in Malay language are "*an*","*i*", "*kan*", "*nya*", "*lah*", and "*kah*". Table 2.5 shows the example the usage of the suffixes.

Table 2.5: Example for Suffix

Root word + Suffix	Words
Bulan + an	*Bulanan*
Milik + i	*Miliki*
Ada + kan	*Adakan*
Kereta + nya	*Keretanya*
Kenapa + kah	*Kenapakah*

2.4.3.3 Prefix – Suffix Pair

Prefix – Suffix pair is the most frequent used in Malay words. The example of common prefix – suffix pair are "*beR-kan*", "*beR-an*", "*di-i*", "*ke –an*", "*ke-an*", "*memper-kan*", and "*se-nya*". Examples for prefix-suffix pair are shown in the Table 2.6 below.

Table 2.6: Example for Prefix-Suffix Pair

Prefix + Rootword + Suffix	Words
beR + teman + kan	*Bertemankan*
Di + adil + i	*Diadili*