CLASSIFYING GOOD AND BAD WEBSITES

Koo Ee Woon

Bachelor of Science with Honours
(Cognitive Science)
2015
DECLARATION OF ORIGINAL WORK

This declaration is made on the 22 day of JUNE year 2015.

Student's Declaration:

I, KOO EE WOON, 36546, FACULTY OF COGNITIVE SCIENCES AND HUMAN DEVELOPMENT, hereby declare that the work entitled, CLASSIFYING GOOD AND BAD WEBSITES is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

22 JUNE 2015

KOO EE WOON (36546)

Supervisor's Declaration:

I, DR. LEE NUNG KION, hereby certify that the work entitled, CLASSIFYING GOOD AND BAD WEBSITES was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a partial/fulfillment for the conferment of BACHELOR OF SCIENCE WITH HONOURS (COGNITIVE SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work.

22 JUNE 2015

(DR. LEE NUNG KION)
I declare this Project/Thesis is classified as (Please tick (✓)):

☐ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1972)*
☐ RESTRICTED (Contains restricted information as specified by the organisation where research was done)*
☒ OPEN ACCESS

I declare this Project/Thesis is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (✓)):

☒ YES
☐ NO

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself/herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature: ___________________________ Supervisor's signature: ___________________________
Date: 22 JUNE 2015 ___________________________ Date: 22 JUNE 2015

Current Address:
257-C, Jln Kenanga 2/2, Kampung Lapan,
Bachang, 75200, Melaka, Malaysia.

Notes: * If the Project/Thesis is CONFIDENTIAL or RESTRICTED, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.
CLASSIFYING GOOD AND BAD WEBSITES

KOO EE WOON

This project is submitted
In partial fulfilment of the requirements for a
Bachelor of Science with Honours
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
(2015)
The project entitled 'Classifying Good and Bad Websites' was prepared by Koo Ee Woon and submitted to the Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours (Cognitive Science)

Received for examination by:

(DR. LEE NUNG KION)

Date:
22nd June 2015

Grade
A
ACKNOWLEDGEMENTS

First of all, I would like to express my earnest sense of appreciation to my supervisor, Dr. Lee Nung Kion, who offered his continuous encouragement and advice throughout the course of this Final Year Project. I thank him sincerely, for his guidance, understanding, wisdom and encouragement and for pushing me farther than I thought I could go. I appreciate the time and effort he invested in supervising and mentoring me throughout this journey.

I also wish to thank all my course mates for the assistance and suggestions given throughout the project. My joy knows no bounds in expressing my cordial gratitude to my friends for helping me survive all the stress from this year and not letting me to give up. I am blessed with precious friendship and spiritual support given by all.

Not to forget, I would like to take this opportunity to express the profound gratitude from my deep heart to my beloved parents, grandparents, and my siblings as well as cousins for their love and continuous support – both spiritually and materially.

I humbly extend my thanks to one and all, who had aided me directly or indirectly in completing this Final Year Project.
# TABLE OF CONTENTS

LIST OF TABLES .......................................................................................................................... v
LIST OF FIGURES ..................................................................................................................... vi
ABSTRACT .................................................................................................................................... vii
ABSTRAK ..................................................................................................................................... viii
CHAPTER 1_INTRODUCTION .................................................................................................... 1
CHAPTER 2_LITERATURE REVIEW ........................................................................................ 5
CHAPTER 3_METHOD ............................................................................................................ 14
CHAPTER 4 RESULTS AND DISCUSSION ............................................................................ 19
CHAPTER 5 CONCLUSION AND RECOMMENDATIONS ....................................................... 24
REFERENCES ......................................................................................................................... 26
APPENDIX A_FEATURES EXTRACTION CODE ..................................................................... 30
APPENDIX B_CLASSIFICATION CODE .................................................................................. 34
LIST OF TABLES

Table 1 HTML Tags Selected for Classification ........................................................................... 17

Table 2 Comparison of Various Classifiers used for Web Page Classification .......................... 20

Table 3 5-Fold Cross Validation with 160 Training Data and 40 Testing Data for Each Fold .... 21
LIST OF FIGURES

Figure 1 Example screenshots of five ugly and five beautiful pages. ...................... 9

Figure 2 Example screenshots of three web pages for each of the four website topic classes. The images shown here are representative of the look and feel in for those website topics. ............ 10

Figure 3 The project framework. ........................................................................ 14

Figure 4 Data of average frequency of occurrence of tags. .................................... 19

Figure 5 5-fold cross-validation accuracy results. ............................................... 21
ABSTRACT

Websites classification has become a vital subject matter as most websites are increasingly being used as a platform for various applications. These web pages often contain semi-structured content which make the classification process challenging. This paper addresses the use of machine learning techniques to classify good and bad websites. The classification process is made easy by using set of features generated from HTML codes. The performance of the 21 features were evaluated by using three machine learning techniques: support vector machine (SVM), naïve bayes, and nearest neighbor classifiers. The good and bad websites were distinguished by the set of features obtained through counting of the HTML tags. A total of 200 websites were collected from machine learning task. The results obtained indicate that the features are useful for classification tasks with average accuracy of 80.50% for SVM classifier, 77.00% for naïve bayes classifier, and 72.50% nearest neighbor classifier. Hence, SVM classifier achieved the highest accuracy among all. This project illustrates that it is possible to classify websites as good or bad by using the underlying tags along with the machine learning algorithms.

Keywords: web classification, machine learning techniques, HTML tags, SVM, Gaussian naïve bayes, k-nearest neighbor, supervised machine learning approach, World Wide Web.
ABSTRAK

Klasifikasi laman web telah menjadi satu perkara penting kerana semakin banyak laman web telah digunakan sebagai platform untuk pelbagai aplikasi. Laman-laman web sering mengandungi kandungan separa berstruktur yang menyebabkan proses klasifikasi lebih mencabar. Kertas kerja ini membincangkan penggunaan teknik pembelajaran mesin untuk mengklasifikasikan laman web kepada baik dan buruk. Proses pengelasan adalah lebih mudah dengan menggunakan set ciri-ciri yang dijana daripada kod HTML. Prestasi 21 ciri dinilai dengan menggunakan tiga teknik pembelajaran mesin: mesin vektor sokongan (SVM), Bayes naif, dan penjodoh jiran terdekat. Laman-laman web yang baik dan buruk dapat dibezakan oleh set ciri-ciri yang diperoleh melalui pengiraan tag HTML. Sebanyak 200 laman web telah dikumpulkan dengan menggunakan tugas pembelajaran mesin. Keputusan yang diperolehi menunjukkan bahawa ciri-ciri yang dipilih adalah berguna untuk kerja-kerja klasifikasi dengan ketepatan purata 80,50% untuk pengelas SVM, 77.00% bagi pengelas Bayes naif, dan 72,50% pengelas jiran terdekat. Oleh itu, ketepatan yang dicapai oleh pengelas SVM adalah tertinggi dalam kalangan semua. Projek ini menggambarkan bahawa ia adalah mungkin untuk mengklasifikasikan laman web kepada baik atau buruk dengan menggunakan tag asas bersama-sama dengan algoritma pembelajaran mesin.

CHAPTER 1
INTRODUCTION

The World Wide Web, which is also known as the Web, is a platform for Internet users. According to Chuan-Chuan Lin and Lu (2000), the usage of internet has been increasing drastically since the preliminary application in 1994. This also reflects that there will be more interaction between users and web-based applications. Information spread quicker and wider due to the connectivity of the Internet that enables users to communicate via one-to-many and many-to-many database relationship (Cheung & Lee, 2005). Thus, many companies take these advantages to set up the Web as a virtual marketplace to communicate with the customers since the Internet became the fastest growing mechanism for electronic industry (Chuan-Chuan Lin & Lu, 2000). Therefore, it is important to know users’ preferences toward the website as it is the prime key interface for e-commerce business, information provision and promotional activities. However, it is still a challenge for all web designers and manager to develop a site that is able to fulfill the user needs (Price as cited in Palmer, 2002). A successful and well-designed web needs to consider the usability and other design criteria so that it can satisfy its users (Nielsen, 2000; Pearrow, 2000).

Background of the Study

Hypertext markup language (HTML) is one of the common languages used in writing web pages and a collection of these pages can form a website. According to the Total Number of Websites (2013), there were a total of 14.3 trillion web pages ‘live’ on the Web and this number will keep growing undoubtedly. New web pages are being designed, developed and published daily by people. This also indicates that there is a vast difference among the designs of the web pages with one another. According to (Lindgaard, Fernandes, Dudek, & Brown, 2006), a
person’s visual appeal towards a web page can be formed within 50 milliseconds. Thus, the most crucial part of a web page is the first impression it gave to the people who visited it (van den Berg, 2010). Web user is able to grasp an idea on the type of the web page by looking at the specific and attractive features of the page even without reading the words in the page.

There are many types of web page classification methods done by various researchers. Compared to text classification, web page classification is more difficult due to the noise contained in the page. However, the quality of information retrieval can be enriched by effective web pages classification (Xu, Yan, Qin, & Zhu, 2011). Earlier research done by de Boer, van Someren, and Lupascu (2008) showed that the possibility of classifying web pages based on the “Look and Feel”. Besides, there was a study on the use of colour and composition to classify web pages (van den Berg, 2010). Another research is carried out to classify the web pages using its genres (Kennedy, 2004). Moreover, Shih and Karger (2004) developed a model and algorithm of machine learning for automating web page classifier tasks by using URL classifier and table-based classifier.

These researches have shown that classifying web pages is a very interesting part of text-classification. Previous researchers have proven that there are many ways of web page classification. Although the extracted features using machine learning are effective for classifying pages, they still could not figure out the exact factors that make web pages look good or bad.

**Problem Statement / Motivation of the Study**

This study focuses on the use of machine learning techniques to classify the good and bad websites. In order to distinguish good or bad websites, this project emphasizes on extracting features by using machine learning that are parallel to how a human perceive a web page to be
good or bad. A good web page initiates the user to revisit it again because the useful information
can be easily accessed and the web page is structured in a way that facilitates access and
navigation (Matera, Rizzo, & Carughi, 2006). Thus, this project approach is to extract features
correlated to the usability and aesthetic view of the web pages’ design.

**Purpose and Objectives of the Study**

The main purpose of the study is:

i. to classify websites as good or bad using machine learning techniques.

The objectives of this study are:

i. to compare three machine learning techniques for classifying websites as good or bad

ii. to extract discriminative features from websites based on usability principles from
websites for classification.

**Research Questions**

1. Which machine learning techniques: Support Vector Machine (SVM), Naïve Bayes and
   Nearest Neighbor perform better in website classification problem?

2. What usability features are useful for discriminating good from bad websites?

**Definition of Terms**

**Websites.** Websites are referred as a gigantic storehouse of information that links people
and provide accessibility to billions of web resources through internet. A text-based web pages
description language called HTML is used to convey all the information (Calongne, 2001). A
well-structured layout of the web encourages users to revisit the pages.

**Web usability.** Usability is a factor that defines the qualities and user-friendliness of a
product from the point of view of its consumers (Matera et al., 2006). When applied to the web,
usability is related to how successful a user is able to perform tasks in pages (Wah, 2011). In brief, the usability of a web page is solely depending on the users.

**Significance of the Study**

Hopefully the results of good and bad websites classification using machine learning based on human perceptions in this study will be able to be improved. Besides, this study is able to help web page designers to design a good web page with the help of automated machine learning techniques. As a result, designing a web page can be less time consuming as these techniques are able to identify the good and the bad features.

**Scope of the Study**

This project focuses mainly on two types of websites, which are, good and bad websites.
CHAPTER 2
LITERATURE REVIEW

Web Quality

How a user determines a good or a bad quality website is depending on their perception towards the web design. Thus, website quality is very subjective as it is affected by users' perception. The web quality that perceived by users is related to the usability of the website and user experiences (Sørum, Andersen, & Clemmensen, 2013). Despite that, web quality is still perceived as a complex concept which the measurement for it is both not easy and not direct (Aladwani & Palvia, 2002).

With the advancement of technology nowadays, web-based applications made are indeed complex and highly sophisticated. Quite a lot of domains have been influenced by these applications by given access to information and services by a variation of users showing diverse characteristics and backgrounds. The usability of a website can determine the success or failure of the website (Matera et al., 2006). This is also why adopting methods for usability evaluation before and after the application deployment has attracted the attention of most industries. Hence, defining the method of measuring the web quality seems to be the main challenge for most of the Web Engineers (Matera et al., 2006). According to the research done by Aladwani and Palvia (2005), website quality is measured by technical adequacy, web content and web appearance. In Sørum et al. (2013) study which focused on the government websites, website quality is mainly related to user-friendliness, effective website usage, content-related issues and accessibility.

Palmer (2002) mentioned that the Human Computer Interaction literature as well as Web-specific usability research is attracted by the usability and design of websites. Usability is commonly recognized as an element of system quality indicating the answers to many
unsatisfying interactions with technology (Matera et al., 2006). The quality of products and systems is described from the perspective of humans who use them.

According to ISO 9241, usability is defined as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”. In this context, effectiveness is referred as the ability for users to achieve certain goals while efficiency is completing a certain goals without wasting any efforts. On the other hand, satisfaction is described as the ease of use that satisfies users.

Other than ISO 9241, Nielsen (1993) believed that usability can be defined as learnability, efficiency, memorability, few errors, and users’ satisfaction. Learnability is the ease of learning to use the web, efficiency refers as the ease to navigate the web after learning it, memorability is the ease of remembering the Web layout and navigation after a period of non-use, few errors indicate easy recover if error occurs and users’ satisfaction measure how users’ feel when browsing the Web. In addition, Nielsin also mentioned that usability of information systems was equivalent to a set of design principles which include consistency of interface such as common placement of navigational tools, response time, mapping and metaphors, interaction styles and multimedia and audiovisual. The study done by De Wulf, Schillewaert, Muylle, and Rangarajan, (2006) shows that designers need to take note of the structure and the websites layout and the uses of multimedia tool to increase users’ interactivity. As the Web grew into an essential interface, Nielnson (2000) extended these design principles which included navigation, response time, credibility and content to focus more specifically into Web environment. Iwaarden et al. (as cite in Sørum et al., 2013) also discovered that fast access (response time) and easy navigation on the website are the important aspects for users to interact with the Web.
Navigation allows users to seek and acquire information freely in a website. Therefore, good links are needed to ensure good navigation. In addition, a well-structured page layout, optimum number of graphical usage and accurate content will provide decent accessibility for the website. A number of designs needed to be taken into consideration to create a maximal interaction between the users and the Web. As a result, users will most probably return to revisit the website if they think the website is usable. In short, a user perceived web-based application is depend on its usability.

It is very common and easy for anyone to come up with a website that functions and looks pleasant at the same time. However, an effective and universally acceptable design needs user’s satisfaction, reduced and low-level learnability, plus able to attain the all the usability goals (Nielsen, 1992). The order of activities being performed however is negligible.

Recognizing the nature of target audience helps the designer in great selection of good desktop publishing and layout considerations, such as the use of color, background images, fonts, styles, and multimedia components. Moreover, it also helps to determine user’s degree of control through the activation of objects and hypertext selections (Champeon, 1999). For an instance, if a site is being designed for Goths whom enjoys a lifestyle that is comprised of pathos, passion and the feeling that the world is a dark place indeed, will certainly not be the same if it is built for Young Republicans. The target users may have their own unique preferences even though they are from similar age group, sex, demographics. For example, in the choice of color, Goth prefers black and dark backgrounds with astounding images and fonts whereas the Young Republicans will most likely include metaphors which foster the notion of patriotism and use red, white and blue fonts and images. User profiles signify the psychological characteristics, knowledge, general experience, and task experience of the respective user community as in a
way of representation (Mayhew, 1992). Similar natures of users are classified together. The
determination, reading level, color sensitivity, culture, primary language, software knowledge, and
background experiences of users is important in changing the designs.

The main goal of the professional web designer is to predict the design of a usable
website that fulfills all the needs (Calongne, 2001). It is highly acclaimed that effective user
interface designs are greatly recommended in future website design to secure the usability and
productivity factors.

Related Works

The method of using colour and composition to classify web pages was proposed by van
den Berg (2010). He developed a number of new feature extractors which are capable of
extracting human-readable features based on the use of colour and composition in a web page.
The composition feature extractors are used to identify the photos and columns in a page while
the colour feature extractors are being used as colour palette extractor and then to analyse this
palette (van den Berg, 2010). This experiment was carried out by using the Naïve Bayes
Classifier and used 10-fold cross validation to check the classification accuracy. The result
showed his method of classification is effective to extract high level design-related features from
web-pages. However, this research was focusing on the visual appearance of the web pages
instead of from all different aspects.

In the research done by de Boer, van Someren, and Lupascu (2008), they classify the web
pages according to the visual features. The results of three different web page classification
experiments were presented. The first two experiments are binary classification tasks.
Experiment one is on aesthetic value which included beautiful and ugly. Figure 1 shows the
example screenshots of beautiful and ugly web pages.
Figure 1. Example screenshots of five ugly and five beautiful pages.

Experiment two is about recency of web pages. As years pass by, web design has been changing in a fast pace. This experiment is to find out whether web pages can be classified into old and new based on the visual classification. The last experiment is different from previous two as it was conducted to classify web pages on the website topics. However, the aim of the experiment is still to classify according to visual appearance. For example, web design blogs have a highly designed visual appearance themselves, while newspaper sites will have a lot of text and images. Figure 2 below shows three representative web pages from this study.
Figure 2. Example screenshots of three web pages for each of the four website topic classes. The images shown here are representative of the look and feel in for those website topics.

The other possible way of classifying web pages is by the genres. This method was proven by Kennedy (2004). He used the web pages' genre for ranking where one genre is known to be more valuable to a particular query than other genre. In the study, features were extracted and used to determine the web page's genre. Three main genres were used, for instance, personal, corporate and organizational home pages (Kennedy, 2004). The feature sets were used to train and test a neural network for genre classification. Firstly, the feature such as linking structure and key terms from a set of web pages are extracted and recoded according to their assigned genre. Then, the extracted features are being inserted to the neural network to test with feed forward neural network with no middle layers. The features selected were able to classify the genres fairly well. On the other hand, when noise was introduced, the results for the precision and recall were reducing significantly. This method is effective for corporate and personal home
pages but not effective for organizational home pages. Organization home pages represent any home page that is not personal or corporate. These pages contain a much diverse set of themes which include religious, society, non-profit organizations and sports. Due to the large range of topics and genres of web pages, many troubles were caused when noisy data were added in. Due to huge diversity of noisy information embedded in web pages, web page classification seemed to be harder than pure text classification. Therefore, Shen, Chen, Yang, Zeng, Zhang, Lu, and Ma (2004), proposed a new web page classification based on web summarization. A new web summarization-based classification algorithm was generated. Result shown that the proposed classification algorithm was able to achieve roughly 8.8% improvement compared to pure text based classification (Shen et al., 2004). This technique only enables to classify text but not images as the images produced are too noisy.

A system is meant to search the relationships between the web page metrics and the page’s ratings, was proposed by Ivory, Sinha and Hearst (2000). This approach was then successfully implied in the field of automated essay grading (KukichK, as cited in Wah, 2011). In addition, Ivory, Sinha and Hearst (2001) also created a tool known WebTANGO that functions by extracting simple, low-level metrics from web pages like number of words and links, average image size and more. WebTANGO only extracts HTML web pages from Internet and stored in web database. The researchers used linear discriminant analysis in order to find correlations between page metrics and page’s classification, besides classifying the web pages into good and not good.

To summarize, this study aim is to apply machine learning techniques on classifying good and bad web pages according to human’s perception to web page classification, rather than
focusing part of content or features of the web pages. This will show that the special nature of Web pages have a large impact on the classification performance.

Features

The following are the features used in the project done by Wah (2011).

Word count. The main components for the user to comprehend on the web pages are words. One of the key elements of many of the metrics that are measured is word counting. The definition chosen was a word that is delimited by any non-alphanumeric character. This has some downsides. The hyphenated words are being totaled as two words or more, for example, “pre-school” and “mother-in-law”. In addition, the quoted words such as “I’m” and “you’ll” will also show up two words.

Text cluster count. The blocks of text that are detached from other text by whitespace are called text clusters. The whitespace is referred as a two-column document which has at least two clusters per page, more if paragraph breaks exist. The likely text clusters by finding blocks of text halted by two or more <br> tags that interrupt unbroken line output. There is no whitespace separator for one <br> tags because there is only one line break.

Total links. Links are the necessary fundamentals of the navigation design. Web pages that have too many links are less likely to contain a rich textual content. This is due the fact that such pages just contain anchors to other pages. Number of links in a web page was calculated using regular expression patterns by searching for the pattern “?href=\".*\"” in the page body. A web page is considered rich, if the count of the contained links exceeds a predefined threshold. A number of usability studies have been carried out to identify the number of links presented on a page (the breadth) and the number of levels (the depth) needed for information searching, and other aspects of navigation structure.
**Total image.** Web designers are encouraged to minimalize the amount of text colors used (Flanders & Willis, 1998). Display text and Body text color measures report the number of unique colors used for body and display text. The measure do not assess if different colors are used for body and display text.

**Total body word and total sentence.** By minimizing the number of words in sentences and reducing the number of sentences in paragraph of a web document, the designers are able to enhance users reading comprehension. Guidelines proposed that a sentence should not exceed twenty words and a paragraph should not have more than six sentences for the readability of prose text,

**ALT, no ALT image count and unsized image count.** Numerous guidelines present that all images should possess an alt tag. The literature on web design guidelines was researched for additional features, similar to the metrics compilation process conducted by Ivory et al (as cited in Wah, 2011).

**Unique image count and average animation count.** Some of the features which are yet associated with design elements by experts, and were easily measured are added as features. This is due to neglecting potential features simply because no expert has proclaimed the efficacy of the features is not advisable. Besides, the current consistencies in guidelines show that there is a need to consider the new features.