Faculty of Information and Communication Technology

THE ANALYSIS OF METADATA BASED CLASSIFICATION FOR CLASSIFYING EDUCATIONAL WEBSITE

Mohd Nazrien Bin Zaraini

Master of Science in Information and Communication Technology

2016

THE ANALYSIS OF METADATA BASED CLASSIFICATION FOR CLASSIFYING EDUCATIONAL WEBSITES

MOHD NAZRIEN BIN ZARAINI

A thesis submitted
In fulfilment of the requirements for the degree of Master
in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2016

DECLARATION

I declare that this thesis entitle "The Analysis of Metadata Based Classification for Classifying Educational Websites" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in the candidature of any other degree.

Signature

Name Mohd Nazrien Bin Zaraini

Date

APPROVAL

I hereby declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality as a partial fulfillment of Master of Science in Information and Communication Technology.

Signature

Supervisor's name Professor Dr. Burairah Hussin

Date

DEDICATION

To you. Yes, you.

ABSTRACT

Initially websites can be easily categorized based on its domain extensions. But due to the explosion of the internet, the domain name restrictions are no longer being adhered. Web classification can help to categorize websites, especially educational websites that being the focus of this research. Classification will be done based on content and metadata in order to get the impact of metadata implementation in terms of classification accuracy. Three sets of 200 pre-determined educational websites taken from DMOZ directory utilized as training data. This is the total number of educational websites with metadata information available in that directory. For content based classification, keywords extracted from the contents and TF-IDF ranking used to get the top educational keywords. These keywords used as a training dataset attribute for educational web classification. The same method goes for metadata based classification, but the difference is that the keywords were taken from its meta description. One class support vector machine method was used because this research is focusing on single class classification only. Cross validation technique and two sets of test data; all educational websites and various categories of website will be used to validate this research. The results shows that content based classification gives more accuracy compare to metadata. Top ranking educational keywords and the analysis of metadata implementation known from this research based on the information retrieval and web classification process.

ABSTRAK

Pada mulanya, mudah untuk mengkelaskan laman web berdasarkan nama domain. Tetapi selepas internet menjadi sangat popular, sekatan nama domain sudah tidak lagi dipatuhi. Proses klasifikasi laman web dapat membantu mengkelaskan laman web, terutamanya web pendidikan yang menjadi fokus dalam kajian ini. Pengkelasan akan dijalankan berdasarkan kandungan dan metadata untuk mengetahui kesan memasukkan metadata pada ketepatan proses pengkelasan. Tiga set data yang mengandungi 200 laman web pendidikan yang telah dikelaskan diambil daripada direktori DMOZ digunakan sebagai data latihan. Ini adalah jumlah web berkaitan dengan pendidikan yang mempunyai kandungan metadata dalam direktori berkenaan. Untuk proses klasifikasi berdasarkan kandungan web, kata kunci diekstrak daripada laman web berkenaan dan TF-IDF digunakan untuk mendapatkan kata kunci pendidikan terbaik. Kata kunci ini digunakan sebagai data latihan bagi mengkelaskan web pendidikan. Kaedah yang sama juga digunakan untuk proses klasifikasi berdasarkan metadata, tetapi perbezaannya adalah data latihan akan diambil dari metadata web berkenaan. Kaedah "one class Support Vector Machine" digunakan kerana kajian ini brtumpu kepada klasifikasi kelas tunggal. Teknik cross validation dan dua set data latihan; semua web pendidikan dan web pelbagai kategori akan digunakan untuk menguji ketepatan klasifikasi. Keputusan menunjukkan bahawa penggelasan menggunakan kadungan mempunyai ketepatan yang lebih baik berbanding matadata. Kata kunci terbaik dan kesan memasukkan metadata dapat diketahui berdasarkan proses pemerolehan informasi dan klasifikasi yang dijalankan.

ACKNOWLEDGEMENT

First and foremost, thank you Allah for giving me the opportunity to continue my studies also giving me the strength and guidance to complete my research and thesis.

I would like to express my sincere gratitude to my supervisor Prof. Dr Burairah Hussin for the continuous support of my master study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research, writing of this thesis and also during my ups and downs. I could not have imagined having a better supervisor and mentor for my master study.

Thank you to all the members of OPTIMASS, BIOCORE and IS3 research group, The Gedabarg and also Anak Didik Prof Bu group for all the memories, tears and joys and for the endless support during my studies.

And also a special thanks goes to Dhana, Shatilah, Andita and Kak Khyrina for all the helps, motivations and the never ending support during my studies. Without them it is almost impossible for me to complete this research and thesis.

And last but not least, I would like to thank my family for believe in me, for the support and encouragement throughout my life.

Thank you all. May Allah repay your kindness a hundredfold.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DF	Document Frequency
DLF	Digital Library Federation
e-GMS	e-Government Metadata Standard
HTML	Hypertext Markup Language
HMM	Hidden Markov Model
IDF	Inverse Document Frequency
IEEE	Institute of Electrical and Electronics Engineers
IR	Information Retrieval
IWPCM	Illicit web page classification method
KNN	Nearest Neighbor
MYREN	My Research and Education Network
METS	The Metadata Encoding and Transmission Standard
NRENs	International Research & Education Networks
NSFNet	The National Science Foundation Network
OCC	One-Class Classification
ODP	Open Directory Project
OSVM	One-Class Support Vector Machine
SVM	Support Vector Machine
TEIN	Trans-Eurasia Information Network

TF	Term Frequency
TLD	Top Level Domain
URL	Uniform Resource Locator
WSDL	Web Service Definition Language
YASS	Yet Another Suffix

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Research Background

Websites can be define as a place, which is where a company or an organization, or a unique individual, insert data that connected to the Internet (Oxford Advanced Learner's Dictionary, 2015a).

It is one of the important sources of reference for educational purpose. Many educational references can be retrieved from the internet easily nowadays. But among them, not all references are useful and informative. Based on the Web Internet Surveys conducted by Netcraft, there are a total of **857,927,160** websites in the world until May 2015 (Netcraft, 2015). Those numbers include all type of website like educational, social media, news and also spam websites. Due to the enormous number of websites it is hard to distinguish between a good education reference website and other type of websites.

In the early days, each website was assigned to a special Top Level Domain (TLD) type based on its category. The definition of TLDs may be vary and depending on what particular definition that needed by certain organization (Leiba, 2009). Table 1.1 below shows the early domain extension and its function.

Table 1.1: Six early domain name extension and its functions

Domain extension	Function
.com	Commonly used for commercial purpose. This is not meant for the Internet's commercial useage as it was heavily discouraged in the past. "dot com" domains are usually for a research department of a company or some division that had a collaboration with a government body or university until the early 1990s. Currently being managed by VeriSign.
.net	Initially, it was being introduced for the usage of domains that point to a shared network of computers, or "umbrella" sites that work as the portal to a group of several smaller websites.
.org	Used for non-commercial organizations that does not belong to any other category. Public Interest Registry managed the .org domain now.
.edu	Specially allocated for educational organizations in the US for example, colleges and universities in which being handled by Educause.
.mil	Specially designed for the military organizations of United State. Department of Defense Network Information Center of United State handles this domain currently.
.gov	For US government organizations, excluding the military. General Services Administration of United State handles this domain currently.

But then in the early 90s, the World Wide Web exploded, causing domain names to explode as well. People are no longer adhere to this restriction as most of the domain name have been taken and it is said that there is no any really significant reason why it should be restricted to only those TLDs (Leiba, 2009). Therefore, in the year 1992 US legislation removed restrictions on the interconnection of commercial traffic with The National Science Foundation Network (NSFNet) (Lenard and White, 2011).

NSFNet was introduced in 1985 to bolster and support advanced networking among education institutions and research departments in the US. National Science Foundation funded the program of coordinated, emerging projects that was financially that acts as the cross country back-bone of the internet. (Mills and Braun, 1988). When the restriction was removed, it causes, the .com other type of TLD domain can be used for any type of website without any restriction.

Due to the removal of TLD restriction, major players in the internet industry like AOL, Yahoo and Google have created manually listed web directory. AOL has a joint partnership with DMOZ directory while Yahoo and Google have their own web directories called Yahoo Directory and Google Directory. The main purpose is to categorize the web based on human judgment and list them accordingly. But since Google stops its directory in 2011 it seems that manual – human based directory are not relevant anymore (Schwartz, 2011). Google believes that is the fastest way to find information needed on the web is by doing web search. However, due to these circumstances, it is believe that identification of educational website is a significant study.

Without any restriction, the educational website started to have an option to choose it domain names. Although most of the websites still follow TLD restriction, but there are still educational websites that chooses to country TLD instead of .edu (Rodgers et al., 2003). In Malaysia for example, the official website of Universiti Sains Malaysia and

Universiti Teknologi Malaysia use .my domain extension instead of .edu.my (Universiti Teknologi Malaysia, 2015), (Universiti Sains Malaysia, 2015).

Educational websites are supposed to be very broad due to the broad definition of educational term itself. Education can be define as a process of training, teaching and learning, particularly in schools or colleges, that can bring benefit on enhancing knowledge and flourish skills (Oxford Advanced Learner's Dictionary, 2015b). Therefore. Educational websites can include many other types of website rather than educational institution website only.

Without the TLD restriction and the irrelevancy of web directories, new methods are needed to identify educational websites.

In order to identify educational website, the process to recognize the characteristic of educational websites need to be done first. For this research, the content and metadata from known educational websites being extracted to know which educational keywords are most commonly used. The educational keywords can be used to characterize educational websites.

Information retrieval (IR) process will be done to pre-determined educational websites. IR process includes several sub-process like document indexing, data extractions, term ranking and text-preprocessing (Baeza-Yates and Ribeiro-Neto, 2011).

Several known educational web sites will be gathered from DMOZ directory for this research. DMOZ is the biggest and most comprehensive web directory that being edited manually by human. It is developed and handled by a passionate, worldwide community of editors voluntarily. At first it was being known as the Open Directory Project (ODP) (About DMOZ, 2015). All of the website listed in DMOZ are based on human judgment and can be used as a training data for web classification.

The data from known education website need to be extracted and will be used as a training data set for classification purpose. Data from other set of unknown website or test website will also be extracted. These data includes the keywords extracted from it web content and also its metadata.

The data from trained and tested website will undergo classification process using machine learning method. If the test sites match with the criteria of the trained website, then it can be classified as educational sites. Each website has their own characteristic based on structure, layout, and content (Stumme et al., 2006). By comparing the characteristic of the test and trained websites, educational websites can be classified.

Classification is arranging a group of data in classes or categories according to shared qualities or characteristics (Oxford Dictionaries, 2012), while web classification can be define as a process of appointing electronic documents based on their content into one or more categories (Yusuf et al., 2010). Other than classification based on its contents, classification can also being done by referring to the metadata implemented on a website (Fathi et al., 2000).

In simple words, metadata can be defined as information about information or data about data. Which is also a structured information that locates, describes, explains, or otherwise make the process to use, make, or manage an information resource becoming much more easier (National Information Standards Organization, 2004). Website's metadata is the text that included in <meta> tag of a website. It often includes information such as web description and keywords.

This study will compare the accuracy of educational website classification based on extracted keywords and metadata to determine the impact of implementing metadata in a website.

1.2 Problem Statement

There are more than 800 million websites on the internet currently; among them there are several different categories of websites such as educational website, entertainment, news, social networking and others. Due to the enormous number of websites, it is quite hard to distinguish between educational website and other type of websites.

Classification approach can be used to solve this problem. There are several approaches of conducting web classification such content based and metadata based which keywords for a website being assigned manually by the web author.

Embedding metadata in a websites is one of the ways to provide structured data on the web (Kinsella et al., 2011). It usually consists of the data that locates, describes and handles a particular resource object, which discovering and obtaining is being aided. The application and advancement of metadata standards are auspicious to the standardize description and sharing of resources. The variety of resources and different requirements of application that prevents resources sharing are the causes for the existing of several type of metadata standards (Li et al., 2008).

These standards include IEEE Learning Object Metadata (LOM), Dublin Core and the Learning Resource Metadata Initiative (LRMI). LRMI is a collaboration of Microsoft Bing, Google, W3C, Yahoo, and Yandex which being launched in 2011 that acts as an extension of schema.org. (Stanković et al., 2014)

Metadata implementation is not compulsory; it is willingly and specially appointed in its usage. Web authors can choose metadata elements and insert them straightforwardly in their site pages voluntarily (Zhang and Dimitroff, 2005). This can cause a problem as web authors tend to put insignificant keywords to their websites to gain more traffic to their websites (Sokvitne, 2003).

There are several research had been done about metadata for example researchers from Yahoo have done a metadata statistic for large web scopus to provide a search-engine centric view on the Web (Mika and Potter, 2012). There are also research being done to study the difference of implementing metadata on a website in the visibility of search engine result (Zhang et al., 2005).

But still there are less or no research being done to study the impact of metadata implementation on web classification especially educational website. Thus that is why this research being done.

There are three main research questions for this study:

1. How to classify a single class educational website with metadata features?
2. How many educational keywords are sufficient enough to be the benchmark or training data for educational web classification
3. How to verify the educational web classification accuracy?

1.3 Objectives

This research is conducted to fulfill the objectives, which are:

- To design the framework to classify the educational website with metadata.
- To analyze the classification method that has been designed.
- To verify the analysis of metadata implementation on educational website classification in term of accuracy.

1.4 Significance of Studies

This research is considered significant as it tends to study the impact of metadata implementation on educational website classification. Due to the removal of TLD restriction and the insignificant of web directories, websites need to be classified

automatically. There are several ways to do web classification including content based and metadata classification (Fathi et al., 2000).

Automatic classification can be done based on its content or the metadata that was implemented into a website (Kinsella et al., 2011). But it is not yet proven that metadata can improve the accuracy of website classification. Although this research only focuses on educational websites, but the methods and approaches that being used can also be applied to another category of websites.

1.5 Research Hypothesis

In order to do web classification process, data being mined and extracted from training websites. The data consist of the web content and also the metadata of the website. In this research a comparison will be made between content based and metadata based classification in order to know which source give the best classification accuracy.

It is believed that metadata can give a better accuracy on educational websites classification as the content of metadata was being insert directly in their site pages (Zhang et al., 2005). Compare to page content, there are more data that are not related to educational content although it is for educational websites.

1.6 Thesis Structure

This thesis divided into five chapters, which describe as follows:

1. **Chapter 1** is the first chapter. It contents problem background, goal, objectives, scopes and significances of this research.
2. **Chapter 2** describes related works in this area as well as related domains which includes the machine learning techniques, the type of classification, information about metadata and such. These would help in understanding this thesis.