A Study of University Website Content Classification Using Machine Learning

H R Mohd Sharul1, I Nor Azman1 and M Mohd Su Elya1

1School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia.

Email: sharulhafiz@utm.my

Abstract. A university website is a gateway to the institution’s information, products, and services. As websites grow into millions in numbers, it is essential to ensure that the content reflects the needs of its students, staff, and other academic institution as their primary users. This research investigates the development of a new framework that uses machine learning techniques based on webometrics and web usability to classify the web pages of academic websites automatically. The framework briefly introduced how it can help classify web content and eliminate unrelated content and reduce storage space. The findings can also be used to analyse other web-based data to give additional insights that may be beneficial for webometrics studies and identify university website’ characteristics.

1. Introduction

A university website is an essential tool in getting prospective students [1-2]. Discovers six university websites users that comprise prospective students, parents, prospective staff, donors, alumni, and events participants. Among the functions are, online learning, product sales or online stores, staff and student information management, news publications, event calendar, centralised notification, staff, and student portals, research repository thesis and publications, and more. Various studies have been done in identifying website targets based on visitors. These are largely carried out through the study of Human-Computer Interaction (HCI) through web usage and accessibility by conducting surveys to several groups of users that involve prospective students, current students, staff, parents, and public visitors. The results of this survey found that the types of websites visited by each group are different. For example, prospective students will find the admission site and the type of courses offered. Current students were visiting the faculty and library site. Staff will enter the official staff portal, and public visitors will look for job vacancies at the university.

Meanwhile, world webometrics rating involving more than ten thousand universities globally has used webometrics method to study website performance based on the qualitative measure of the number of websites, the number of external and internal links, and several files on the website. This study uses the Google search engine as the primary tool to obtain data [3,4]. Studies involving website content have been done extensively using machine learning methods, especially for website classification [5–9]. Machine learning is used because of the enormous amount of data that makes it challenging to manually do content classification work. This machine learning will be taught based on specific algorithms and methods to recognise each data entered. We group these three types of studies: web usability and experience, webometrics, and content analysis, which will be described in the next section. Past studies prove that although the research conducted is from different angles, it has a similar goal: to identify,
analyse, and study the university website's content using various techniques. Table 1 contains brief findings from research techniques from HCI, Webometrics, and Content Analysis. The combination of findings from these three areas of study will provide a link between these research areas starting from the survey study conducted by the HCI group and content analysis that obtains the group of users and the university website's main content. The main groups identified were prospective students, parents, prospective staff, donors, alumni, and events participants [2]. The university website's main content is a list of courses, admission, course schedule and materials, academic calendar, application and registration, and online lectures [10]. However, this list still needs to be improved and refined to get a more uniform list with other HCI studies' results.

Table 1. Findings of research done by three fields of research in university web pages.

| Research area            | Findings                                                                                                                                                                                                 |
|-------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Web usability & experience | There are six university websites users which comprise of prospective students, parents, prospective staff, donors, alumni, and events participants [2,11]. Quick links to the university catalogue, admissions office, calendars, clubs and organisations, news, campus events, library databases, and a virtual campus tour on the university website. List of courses, admission, billing procedure, course schedule and materials, academic calendar, results, exams, application and registration, online lectures [10] |
| Webometrics             | Improving webometric ranking in web size requires the increase of [12]: - updated staff profiles - social and academic activities - research activities - journal repositories - news & notice                                                                                      |
| Content analysis        | Classifying pages and links manually into research and non-research [13]. Using machine learning to classify academic pages into 8 types which are about, business and innovation, discussion, support, research, staff, student life and study [5] |

In terms of webometrics, there are several major studies conducted. Among them is a study to determine why a website was created and why links are made to other sites [7]. Among the findings obtained is a website created to increase webometrics ranking in the World Webometrics Ranking held twice a year by Cybermetrics Lab [12]. To identify the purpose of link creation, [7] found too broad why such links were created, however, most of what was agreed to cover student learning materials, student and staff support systems, and non-academic content, that was done manually on 414 pages. To reduce human subjective opinions during website classification and classification on many websites, automatic classification methods can be done using machine learning. Among the popular techniques used for classification, purposes are Support Vector Machines (SVM) [14,15] and Random Forest (RF) [16–18]. Models developed using trained databases that are partially labelled by humans. Has labelled 2,549 pages individually and this tends towards his opinion which can, of course, be refuted by other researchers [5].

To get a more realistic classification with the university website's current content, this paper would like to propose a framework that combines the field of content analysis, webometrics, and machine learning that can separate the content of university websites that are academic and non-academic.

2. Method
In this study, we prepared a similar method with an investigation [19]. A checklist based on properties or type of content found in the universities’ home page in QS World Ranking. We scanned the universities’ main homepage thoroughly in January 2020 and categorised the properties into two
variables ‘available’ (Y) and ‘Not available’ (N). The collected data have been presented in a table and calculated for this study. Based on the homepage features, we try to classify the website’s content at a deeper level. These require a crawler to gather all the websites before they can be processed, analysed, and the data collected can be used as a dataset be used in machine learning. The classes of the classifications are based on the features found on the homepage.

3. Results and Discussions

There are nine content types and features found on the university web site's homepage, including news, events, repositories, library, faculty, research, contact, give and eLearning. Table 2 shows website features and content types of Top 10 universities in QS World University Ranking. Other features that are not found on the homepage can be found when searching more in-depth into the website. As the top universities’ website is enormous in size and taking much time to crawl, we decided to run this study on one of the local universities in Malaysia, University Teknologi Malaysia (UTM) to get an initial view of the content. The pages are then classified by subdomains as this can roughly indicate the overall content type of a website as in Figure 1.

| University                        | News | Events | Repositories | Library | Faculty | Research | Contact | Give | eLearning |
|-----------------------------------|------|--------|--------------|---------|---------|----------|---------|------|----------|
| University of Oxford              | Y    | Y      | N            | Y       | Y       | Y        | Y       | Y    | N        |
| University of Cambridge           | Y    | N      | N            | Y       | Y       | Y        | Y       | Y    | N        |
| Stanford University               | Y    | Y      | N            | Y       | Y       | Y        | Y       | N    | Y        |
| Massachusetts Inst. of Tech       | Y    | Y      | N            | N       | N       | Y        | Y       | N    | Y        |
| California Inst. of Tech          | Y    | Y      | Y            | Y       | Y       | Y        | Y       | Y    | Y        |
| Harvard University                | Y    | Y      | N            | Y       | Y       | Y        | Y       | Y    | Y        |
| Princeton University              | Y    | Y      | N            | Y       | Y       | Y        | Y       | Y    | N        |
| Yale University                   | Y    | Y      | Y            | Y       | N       | Y        | Y       | Y    | Y        |
| Imperial College London           | Y    | Y      | Y            | Y       | Y       | Y        | Y       | Y    | Y        |
| University of Chicago             | Y    | Y      | N            | Y       | Y       | Y        | Y       | Y    | N        |

Table 2. Website features of Top 10 universities in QS World University Ranking

Framework for university website classification

From the literature review, we can conclude that the relationship between these fields of study can create a basic framework and step-by-step implementation of the study to separate academic and non-academic content on a university website (see Figure 2). This framework begins by conducting a web usability study in the form of a survey to the university website's primary users in identifying what information the user is looking for.

This list of information types will be used as an essential guide to the classification type to be given to the dataset to be labelled. Information from the web usability study will also support webometrics research on the links available on the website to see the relationship between the site and the links generated. This link will also be labelled to strengthen the accuracy of the dataset label further.

The combination of this web usability and webometrics study’s findings will produce a dataset that has been labelled, which will be used as input in the production of a machine learning model. This model will be trained and tuned with the highest efficiency and accuracy possible to classify the dataset.
Figure 1. Initial results of personalize web crawler and classification.

This framework's result is the automatic classification of university website content that separates academic and non-academic content.

Figure 2. University Websites Content Classification Framework

4. Conclusions

The main purpose of this paper is to propose the use of web usability research, webometrics, machine learning and content analysis to automatically classify the content of university websites and try to see other perspectives that can be used to make this classification. To achieve this goal, a combination of these fields of study was created and produced a framework that described the steps to obtain input from one field of study and submitted to another field. Past studies in using machine learning to classify website content using data obtained by self-developed crawlers and data are labelled individually or in small groups. In this research, we found that the size of content in subdomain in initial findings does reflects the type of content found in the university website homepage. Future works involving classification of university websites, the perspective or opinion of the university website's main users should be considered, and this framework is expected to help in this type of research.
References

[1] Hall H and Witek L 2016 *Procedia Economics and Finance* **39** 206–11
[2] Pierce K 2005 *Web site usability report for Harvard university* (Capella University)
[3] Aguillo I F, Ortega J L and Fernández M 2008 *Higher Education in Europe* **33** 233–44
[4] Thelwall M 2008. *J. Inf. Sci. Eng.* **34** 605–21
[5] Kenekayoro P, Buckley K and Thelwall M 2014 types *Scientometrics* **101** 1015–26
[6] Gaifulina D and Chechulin A 2020 *13th International Symposium on Intelligent Distributed Computing, IDC 2019* vol 868, ed Kotenko I., Desnitsky V., Badica C., El Baz D. and Ivanovic M. (Springer) 278–84
[7] Wilkinson D, Harries G, Thelwall M and Price L 2003 *J. Inf. Sci. Eng.* **29** 49–56
[8] Trung N M, Tam N D and Phuong N H 2011 *Proceedings of the Second Symposium on Information and Communication Technology* SoICT ’11 (New York, NY, USA: Association for Computing Machinery) 106–11
[9] Tsukada M, Washio T and Motoda H 2001 *Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development* WI ’01 (Berlin, Heidelberg: Springer-Verlag) 303–13
[10] Kaul S 2006 *Higher Education in India - Seizing The Opportunity* (East Asian Bureau of Economic Research)
[11] Christoun S, Aubin H, Hannon C and Wolk R M 2006 *Inf. Systems Education Journal* **4**
[12] Anyira I E and Idubor I 2020 *Library Philosophy and Practice*
[13] Thelwall M 2001 *J. Am. Soc. Inf. Sci. Technol.* **52** 1157–68
[14] Dandeniya D 2018 *18th Int. Conf. on Advances in ICT for Emerging Regions (ICTer)* (ieeexplore.ieee.org)196–202
[15] Abooraig R, Al-Zu’bi S, Kanan T, Hawashin B, Al Ayoub M and Hmeidi I 2018 *Digital Investigation* **25** 24–41
[16] Qi X and Davison B D 2009 *ACM Comput. Surv.* **41** 1–31
[17] Subasi A, Molah E, Almkallawi F and Chaudhery T J 2017 *Int. Conf. on Electrical and Computing Technologies and Applications (ICECTA)* pp 1–5
[18] Akinyelu A and Adewumi A O 2014 *J. Appl. Math.* **2014**
[19] Rahman A and Batcha M S 2020 *J. of Library & Information Technology* **40**

Acknowledgement

The authors wish to thank the Ministry of Higher Education, Malaysia, for funding this study under the Fundamental Research Grant Scheme (FRGS/1/2019/ICT04/UTM/02/8)