5-28-2020

A Systematic Review of Web Usage Mining Techniques and Future Research Options

Brent Van Aartsen
brent.vanaartsen@dsu.edu

Omar El-Gayar

Cherie Noteboom

A Systematic Review of Web Usage Mining Techniques and Future Research Options

Brent Van Aartsen
Dakota State University
brent.vanaartsen@dsu.edu

Omar F. El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Cherie Noteboom
Dakota State University
cherie.noteboom@dsu.edu

ABSTRACT

Through this systematic review, we review the research of web usage mining (WUM) techniques from 2014 - 2019 in order to understand the current state of WUM research and answer our research questions; (RQ1) what data sources are used in web usage mining, (RQ2) what data analysis methods are used to extract the knowledge, (RQ3) what are the applications of Web usage mining, and (RQ4) what future research can be done in the web usage mining area? Using a PRISMA approach to narrow the initial 778 search results, we completed a full analysis of 74 unique articles from four prominent IS databases. The article analysis showed WUM research is on the decline and revealed Personalization and Recommender Systems are the two most heavily researched WUM applications.

Keywords

web usage mining, WUM, techniques, systematic review

INTRODUCTION

In this ever increasingly Internet connected world, the amount of new data being created every day is immense (Jain, Rawat, Bhandari, 2017). Harnessing that data is a big challenge. One way to begin harnessing that data is with web mining. Web mining is a specific subset of data mining that focuses specifically on web data such as page content, web logs, and navigation structures (Jain et al., 2017). Web mining is comprised of three components, web structure mining, web content mining, and web usage mining. We will focus our attention to the web usage mining.

Web usage mining, (WUM), is an application of data mining techniques on web log data in order to understand the who, what, why, and how of those using a website (Suharjito, Diana, and Herianto, 2016). Throughout this paper we explore the research of WUM techniques from 2014 - 2019 to better understand the current state of WUM research.

Research Questions

The primary driver of this research is identifying the different data analysis techniques utilized by WUM and to identifying how those techniques are applied. A secondary objective is to identify any gaps in the current research to assist in building a path for future research. Our research questions reflecting these goals are:

RQ1. What data sources are used in Web usage mining?

RQ2. What data analysis methods are used to extract the knowledge?

RQ3. What are the applications of Web usage mining?

RQ4. What future research can be done in the Web usage mining area?

CLASSIFICATION FRAMEWORK

To aid in our systematic review (SR), we defined the WUM Classification to Possible Applications Framework (WUMCPA). The WUMCPA defines which components of WUM relate to different potential applications (see Figure 1). Each WUM component can be applied separately by an application or the components can be utilized in combination with one another.

Web Usage Mining

Clustering

In web usage mining, clustering is the process of grouping like users or like pages together. Clusters of users are comprised of users with similar browsing patterns. Page clusters are comprised of pages containing similar content. Clustering algorithms fall in one of the three categories: partitioning methods, hierarchical methods, or model-based methods. Partitioning methods group data based on the need of the user using algorithms such as K-means clustering or Page Gather. Hierarchical methods create structured subgroup clusters using algorithms like BIRCH. Model-based methods find the best fit of datasets and mathematical sets using algorithms like ITERATE and Self-Organizing. (Jain et al., 2017)

Association Rule

Association rules determine if a relationship exists between pages accessed on a website (Dhandi and Chakrawarti, 2016). Each association rule has a basic support, it's popularity, and confidence, it's likelihood of occurring, metrics (Jain et al., 2017). The Markov Model, FP-growth, and Latest Substring Association are examples of algorithms for association mining.

Classification

Classification maps data item sets into pre-defined classes. Each class has its own characteristics defined by the choice of features and extracted features (Jain et al., 2017). Supervised inductive learning algorithms map the data items. HCV, CDL4, and Rough Set Theory are algorithms routinely used in classification (Jain et al., 2017).

Possible Applications

Web Page Prediction

Web page prediction is a way to predict what web pages a user will visit next based on previously visited pages. Web page prediction allows web administrators to pre-cache web pages before a user requests them, allowing the page to load faster. Web page prediction can also reduce the number of unwanted pages a user has to visit to find the content they are looking for. PageRank and Markov Model are two algorithms used for web page prediction. (Chatterjee, Ray, and Bag, 2017)

Personalization

Personalization uses WUM outcomes to display content based on users' interests (Suadaa, 2014). Personalization can also be based on other factors like geolocation or time of day. An example of geolocation-based personalization is having advertisements for stores near your location appear in your browser. To expand on that example, depending on the time of day, you may receive advertisements for nearby restaurants.

Recommender System

Much like personalization, recommender systems work to show users recommendations based on their previous web browsing history and other variables (Herath and Jayaratne, 2017). Recommender systems differ from personalization because the user is requesting the information being provided. For example, searching for "restaurants" on your phone will combine your past history of searching, reviewing, and/or visiting restaurants with your search terms and GPS location to return you restaurants that you may like near you.

Adaptive Websites

Adaptive websites adapt based on the outcomes of web usage mining (Herath and Jayaratne, 2017). Modifications to website navigation, usability, or accessibility are a few examples of items a website may adjusted. The goal of an adaptive website is to respond to users' needs and wants of the site, so the site runs more effectively for the users (Bhargav and Bhargav, 2014).

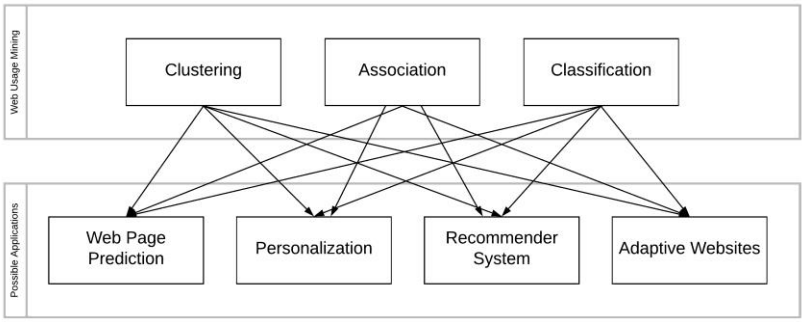

Figure 1. WUM Classification to Possible Applications Framework (WUMCPA)

METHODOLOGY

To complete our SR effectively, we first defined our research questions to direct our SR. We then developed a set of keywords for searching for articles. With our research questions and search keywords defined, we adopted the PRISMA Model (Kadi, Idri, and Fernandez-Aleman, 2017) to help structure the SR. Finally, we applied the WUMCPA to classify and tag articles selected for full analysis.

Search and Article Selection Strategy

Our article searches focused on the keywords (i) "web usage mining", (ii) "WUM", and (iii) "web usage AND mining". The search was performed using Web of Science, ProQuest, ScienceDirect, and IEEE Xplore. Searches were performed on the titles and abstracts only, except in Web of Science where the search was completed on the search field Topic as Web of Science does not allow for searching titles and abstracts only. The Web of Science Topic search field consists of title, abstract, author keywords, Keywords Plus®. The searches were limited to only include papers from 2014 through early 2019 to ensure the research remain focused on the latest methodologies. Additional exclusion factors limited the results to only include journal articles and conference proceedings, in English, and where the full text was available.

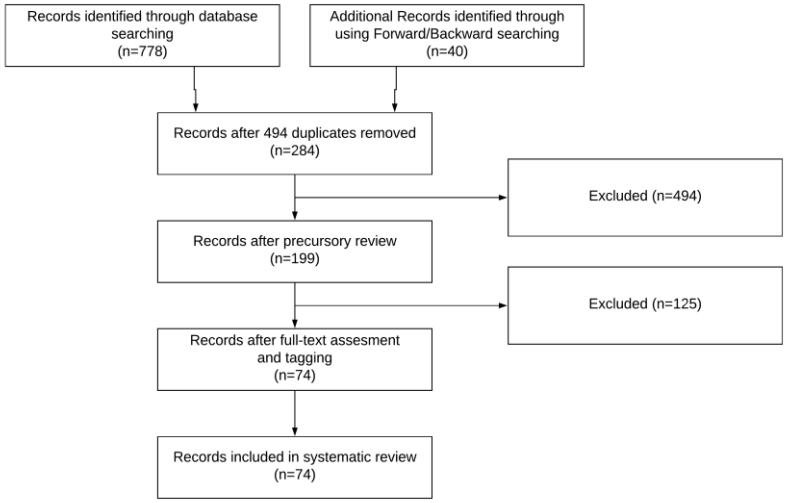

Figure 2. WUM Selection Process using PRISMA Model

Following a PRISMA based approach (see "Figure 2"), we reduced the initial 778 search results by removing all duplicates, leaving 284 articles to investigate (Kadi et al., 2017). Excluding articles without a focus on techniques, we were left with 74 articles to complete a full analysis of. Utilizing the WUMCPA framework, we classified and tagged each of the remaining articles. Completion of the full analysis found all studies to be relevant to our review.

RESULTS AND DISCUSSION

RQ1 - WUM data sources

Our analysis revealed the most prevalent source is the server logs (see Figure 3), which is expected as the primary source of logs for WUM is server logs (Manchanda and Gupta, 2018). Usage of client logs and proxy logs are both about equally prevalent. No matter the year, client logs and proxy logs are each used less than 20% of the time. Client logs can be more difficult to obtain as they require adding cookies or the use of JavaScript on the client machine (Jain et al., 2017).

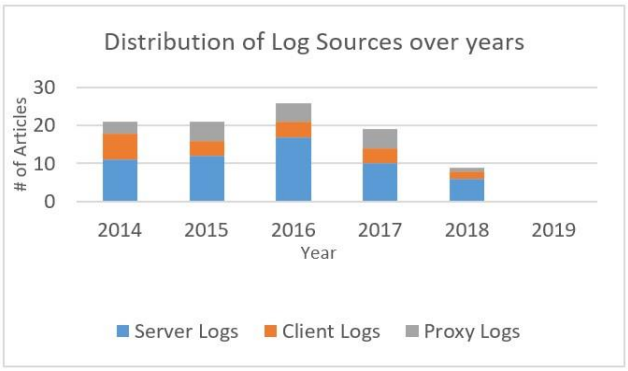

Figure 3. Distribution of Log Sources over years

RQ2 - WUM data analysis methods

We found the two most used algorithms for association are Apriori and Markov Model. Our research revealed 19 clustering algorithms, of which, K-means is the most used algorithm. Finally, of the classification algorithms explored from 2014-2019, K-nearest neighbor and Naïve Bayesian are the two most prevalent algorithms used.

RQ3 - Applications of WUM

Given the emphasis on delivering personalized services to customers, the results here are surprising. More articles have utilized or reviewed recommender systems versus personalization (see Figure 4). In fact, of the 7 articles published in 2018, there were no efforts made on either web page prediction or personalization.

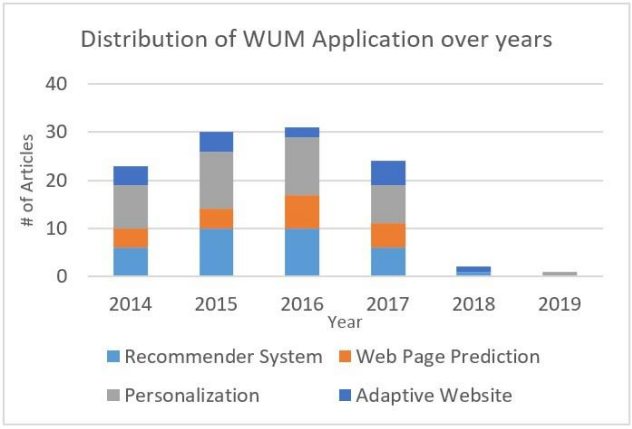

Figure 4. Distribution of WUM Application over years

RQ4 - Future Research

One vantage point for answering the question is to review what the authors of the articles in the study put forth as future research avenues. Another vantage point is to review the data analysis for gaps in the literature we can further investigate. We will discuss both vantage points through the remainder of this section.

Author Suggested Future Research

In nearly every article, authors suggested one or more future research paths. One area several articles discussed was scaling their experiments to larger datasets. In many instances, the test datasets are no longer representative of today's web logs. Repeating the studies using current web log data could reveal new insights. Several articles also discussed applying various methods and techniques to web security by using WUM to discover anomalous traffic (Atta-ur-Rahman, Alrashed, and Abraham, 2017).

Future Research from Literature Gaps

After analyzing the data from our SR, we find several areas where the literature may warrant further research. These areas include additional research in the use of proxy and/or client log sources, the application of WUM techniques for web page prediction, the use of WUM techniques for adaptive websites, and expanding WUM into other subject areas such as online privacy and web security. We will briefly look at each area and how to further the research.

Proxy logs and client logs are both sources that can be hard to obtain as the logs are often not on devices you have access to (Kaur and Aggarwal, 2017). Our analysis shows proxy logs and client logs were the least used data sources. Finding ways to obtain those logs or mitigate the effect of proxies or client caching is one path to explore. Performing a SR of the topic could well lead you to other technology disciplines that have overcome the same challenge.

Our data analysis also revealed the research is least focused on applying WUM techniques to web page prediction or adaptive websites. Further research can be completed to determine which WUM technique is best suited for web page prediction and for adaptive websites. Possible implementations of the research could be increasing the efficiency of content delivery networks (CDNs) or improved performance of software-as-a-service applications that rely on regional datacenters to distribute the applications.

The last potential research opportunity we will discuss is combining WUM techniques with different technology areas. The areas of privacy and web security have already been suggested by the authors of the papers in our reviews. An additional SR could be performed to look for all topic areas making use of web server logs, proxy logs, or client logs. The review could focus on log collection techniques, log processing techniques, scalability, or a host of other issues we have found in WUM.

CONCLUSION

In this paper, we examined the state of research of web usage mining techniques to better understand what research has been done and opportunities for further research. To do so, we used a PRISMA based approach to structure our SR. We also defined the WUMCPA framework in order to classify and tag the articles we reviewed.

The SR analysis shows the algorithms behind WUM techniques have been explored extensively. We proposed several different avenues for expanding the research related to WUM. To recap, those avenues are: repeating past studies using current web log data, performing a SR focusing on mitigating the effects of proxy/client logs, additional research on applying WUM techniques to web page prediction and/or adaptive websites, and performing a SR of technology areas using web server logs to find overlapping areas.

Limitations to this research could also be expanded upon in future research. Most notably, the limitation in the number of databases searched. Though the four databases searched are premier databases, searches in additional databases could be completed to expand the amount of information gathered. Another limitation to the research is the limitation of the keywords used to search. Additional keywords to explore include "web log mining", "web log AND mining" or "web log analysis".

REFERENCES

1. Atta-ur-Rahman, Alrashed, S. A., & Abraham, A. (2017). User Behaviour Classification and Prediction Using Fuzzy Rule Based System and Linear Regression. *JOURNAL OF INFORMATION ASSURANCE AND SECURITY*, *12*(3), 86–93.
2. Bhargav, A., & Bhargav, M. (2014). Pattern discovery and users classification through web usage mining. *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 632–636.
3. Chatterjee, R. P., Ray, C., & Bag, R. (2017). A Comparative Study on Latest Substring Association Rule Mining and Hidden Markov Model. *2017 International Conference on Computer, Electrical Communication Engineering (ICCECE)*, 1–5.

4. Dhandi, M., & Chakrawarti, R. K. (2016). A comprehensive study of web usage mining. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1–5.
5. Herath, D., & Jayaratne, L. (2017). A personalized web content recommendation system for E-learners in E-learning environment. *2017 National Information Technology Conference (NITC)*, 89–95.
6. Jain, S., Rawat, R., & Bhandari, B. (2017). A survey paper on techniques and applications of web usage mining. *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)*, 1–6.
7. Kadi, I., Idri, A., & Fernandez-Aleman, J. L. (2017). Knowledge discovery in cardiology: A systematic literature review. *International Journal of Medical Informatics*, *97*, 12–32.
8. Kaur, N., & Aggarwal, H. (2017). A Novel Semantically-Time-Referrer based Approach of Web Usage Mining for Improved Sessionization in Pre-Processing of Web Log. *International Journal of Advanced Computer Science and Applications*, *8*(1).
9. Manchanda, M., & Gupta, N. (2018). Web Usage Mining: Dynamic Methodology to Preprocessing Web Logs. *HELIX*, 8(5), 3810–3815.
10. Suadaa, L. H. (2014). A survey on web usage mining techniques and applications. *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, 39–43.
11. Suharjito, Diana, & Herianto. (2016). Implementation of classification technique in web usage mining of banking company. *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 211–218.