Journal of Recent Innovations in Computer Science and Technology

A Review of Web Usage Mining Methodology and its Practical Implementation

Anurag Shrivastava [1], ***Bhavana Shrivastava*** [*2]

[1,2]School of Computing, DIT University, Dehradun, India.

***Corresponding Author Email**: bhavana.shrivastava@dituniversity.edu.in

Abstract

Data mining techniques are employed in online usage mining to evaluate user behavior on web pages and uncover usage patterns. There has been a dramatic uptick in interest in web usage mining from academics and industry professionals alike. The objective is to better comprehend how people use websites and cater to the requirements of web-based applications. Extensive research has already been conducted in web usage mining, a substantial and quickly developing subfield of web mining. To put internet usage mining into practice, a clearly defined technique is now required. This article presents a methodology derived from suggestions from literature initiatives. This review aims to help readers gain a firmer grasp of the technical state of Web Usage Mining Methodology and future direction. Researchers and practitioners can use this review to guide better decision-making by connecting theoretical knowledge with actual implementation. Also included a current overview of the literature. A quick review of web usage mining is provided.

Keywords: Web Mining, Web Usage, Data Mining

1. INTRODUCTION

Over the past few years, there has been an extraordinary increase in the accessibility of online information. This has transformed the World Wide Web (WWW) into a massive repository of information covering various fields of interest. Within this data repository, only a few components consist of organized data, primarily sourced from the organizational online transaction processing (OLTP) system [1].

The rapid expansion of the Internet across the world has been the primary catalyst for the growing recognition of cross-lingual text retrieval (CLTR) in recent times. Relevant material is available in several languages. Users may desire to locate documents written in languages other than the one used to generate their query. Out all the numerous newly established CLTR approaches, query translation has been the subject of the most thorough research [2]. CLTR techniques are primarily designed to enable term-based lexical transfer between a specific source and target language pair. However, relying solely on bilingual lexical transfer is inadequate for completely meeting the user's requirement for multilingual Information for Customer Relationship Management is a focal point in the current financial services industry. In today's highly competitive market, establishing a solid relationship between a provider and each of its consumers has become increasingly crucial. However, we encounter obstacles to direct client interaction, such as globalized markets, corporate mergers, and the growing use of remote transaction services like Internet and phone banking [3].

Our study has made a valuable contribution to the field of Web intelligence by providing insights that can be used to advance research in the creation of multilingual search engines and Web directories. The utilization of a multilingual text mining approach has proposed a compelling and innovative avenue for uncovering valuable knowledge, which can be beneficial in the development of multilingual text management systems. Specifically, our method of analyzing text in many languages to uncover linguistic information automatically contributes to CLTR by offering a more cost-effective alternative to the expensive process of manually creating linguistic resources [4] [5].

To make good use of this data and learn new things by creating and using web-based solutions, we need to follow methods and processes, especially since the amount of data being generated is growing at an exponential rate. Web mining is now a whole new area of study with useful tools for both users and webmasters. There are three different types of web mining:

1. Web usage mining is a method for extracting valuable information and patterns from people's online behavior.

2. The term "web structure mining" refers to the practice of analyzing website architecture and design for valuable data.

3. Extracting valuable information from online page text is known as web content mining.

The site can be built using web usage mining approaches such as clustering, sequential pattern analysis, and link rules. Although there is a vast array of uses for web mining, e-commerce and CRM are among the most prominent. These subjects have been the subject of several written works. At the moment, a lot of research is being done in the area of web mining to find ways to get around these problems. It would be impossible to explain all of the methods and uses of web mining in this piece, but the goal is to give the reader a general idea of what it is and point them in the direction of resources that interest them [11] [12].

When web mining, problems can arise, such as wrong or incomplete data, a lack of available tools, the need for customized tools, insufficient sharing of necessary resources, and management-related issues.

1.1. **Overview of Web Usage Mining**

The term "web usage mining" refers to the process of gleaning useful information from server logs by analyzing user browsing and access habits. Commercial websites can greatly benefit from this information in order to increase customer happiness.Web mining is an approach to data mining that aims to discover and extract interesting and useful patterns from large datasets using typical data mining methods. Datasets for web mining are what big data is all about.The primary focus of online use mining is the analysis of data pertaining to user activities as they navigate web apps and websites. Web mining is the practice of discovering useful patterns in data generated by the web and extracting them. It includes web structure mining, web usage mining, and online content mining as its primary sub-areas.

A subset of web mining known as "web usage mining" seeks to automatically detect trends in how users interact with various web servers. There has to be a rethinking of the traditional strategies and methods employed for market research in light of the growing reliance of businesses on the Internet and the WWW for operational purposes. In the course of their everyday activities, many organizations create and amass vast quantities of data. [13] [14].

Web usage mining refers to the process of automatically identifying user access patterns from one or more Web servers as part of Web mining. With the increasing dependence of enterprises on the Internet and the World Wide Web for commercial operations, it is necessary to reevaluate the conventional tactics and approaches used for market analysis in this digital setting. Organizations frequently produce and accumulate substantial amounts of data as part of their day-to-day activities [6].

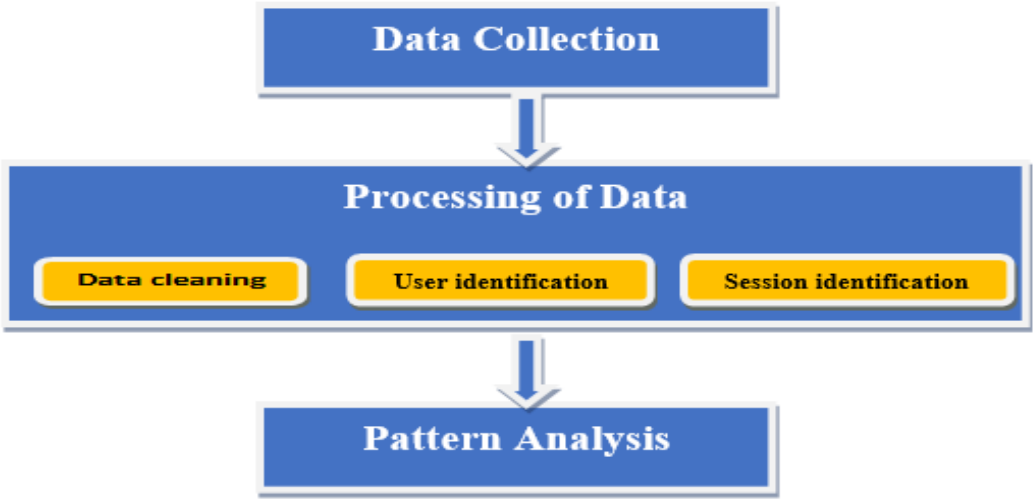

Figure 1. Phases of web mining

1.2. **A methodology for web usage mining**

An important role in website analysis is to examine its usage data, since it offers valuable insights into the site's organization and its effectiveness in meeting user requirements. This knowledge is particularly intriguing for business applications. Within this particular framework, the examination of such data can assist firms in various ways, including the development of cross-marketing strategies and the evaluation of promotional campaign efficacy. Consequently, we established a methodology for identifying the characteristics of visiting groups.

1.3 **Crossed Clustering in Web Usage Mining**

Examining usage data plays a crucial part in website analysis as it provides vital insights into the site's structure and its ability to meet user needs effectively. This information is especially fascinating for business purposes. Within this specific context, analyzing this data can help companies in multiple ways, such as creating cross-marketing plans and assessing the effectiveness of promotional campaigns. As a result, we developed a systematic approach for determining the attributes of groups that visit.

1.4 **P2P Usage Mining**

Given the abundance of information sources accessible on the Internet, Peer-to-Peer (P2P) systems present a unique system design that caters to the large-scale community by offering applications for file sharing, distributed file systems, distributed computation, messaging, and real-time communication. P2P applications offer a robust framework for doing data and compute-intensive tasks, such as data mining [7].

We believe that the interconnected nodes have the ability to collaborate with a designated peer (referred to as a "meter peer") in order to offer the end user a reliable estimation of the patterns present in this extensive distributed database. To assess our methodology, we developed a simulator with the ability to execute a simulated decentralized peer-to-peer system without a predefined structure. Real datasets were utilized in additional experiments.

1.5 **XML Document Mining**

XML documents are more prevalent due to their versatile and adaptable format that may be utilized for many applications. Conventional techniques have been employed to categorize XML documents, breaking them down into their textual components. These methods fail to utilize the inherent organization and content of XML documents [8].

Utilizing Web Page Mining to Enhance Search Engines

Our objective was to establish and assess novel criteria for ranking Web pages, focusing on the presentation of the page. We utilized a test collection of Web sites obtained from various search engines as a result of a certain set of queries. The pre-processing and clustering tasks enabled us to formulate methodological proposals for addressing developing sub-problems, including [15] [16]:

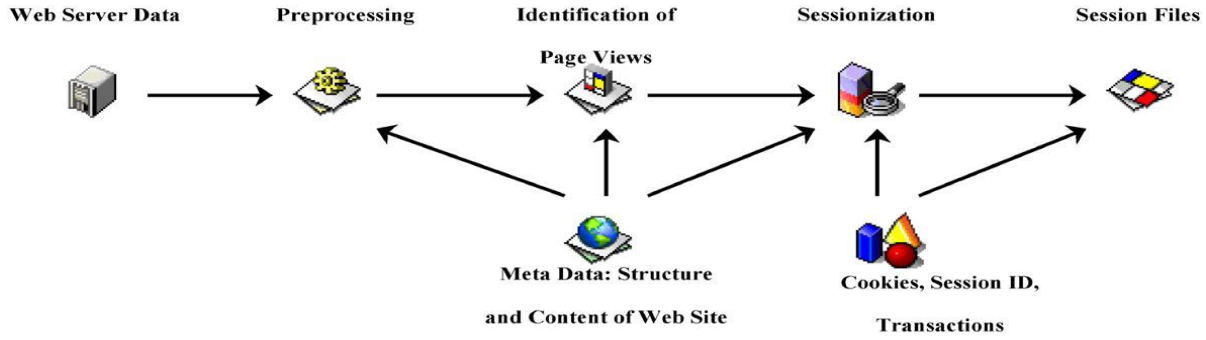

Figure 2: Preprocessing and transformation of web data.

2 Application of the methodology

Our primary objective is to extract structured data from web pages, specifically focusing on information such as products and search results. By extracting such data, it becomes possible to offer services. The course covers two primary methodologies: machine learning and automatic extraction.

Web information integration and schema matching include addressing the challenge of reconciling the diverse representations of the same information found on different websites or even within individual web pages. Identifying or matching semantically comparable material is a crucial issue with numerous practical applications. An analysis is conducted on some established methodologies and challenges.

- Extraction of opinions from online sources: Numerous internet sources provide opinions, such as consumer reviews of items, forums, blogs, and chat rooms. Extracting opinions, particularly those of consumers, is crucial for gathering marketing knowledge and evaluating product performance against competitors. We will provide some tasks and methodologies for extracting information from these sources.

- Knowledge synthesis is facilitated by the use of concept hierarchies or ontologies, which have proven to be valuable in various applications. Nevertheless, manually creating them is exceedingly time-consuming. This presentation will showcase several established techniques that investigate the information redundancy of the World Wide Web. The primary purpose is to amalgamate and arrange the fragments of data found on the Internet to provide the user with a cohesive understanding of the subject matter.

- Web page segmentation and noise detection: In numerous web applications, the primary objective is to extract the core content of a web page while disregarding adverts, navigation links, and copyright notices. Automatically segmenting a web page to extract its key material is an intriguing problem. Several intriguing methodologies have been suggested in recent years.

3 Literature Survey

By utilizing the Naive Bayes method, web data miners can sift through mountains of textual data in search of actionable patterns, insights, and information. Classification and categorization jobs involving text are where it shines. Structure mining, content mining, and use mining are the three primary subfields of Internet data mining that make use of the technique [8].

This article examines the potential advantages of incorporating Semantic Web technologies into healthcare information systems, including enhanced knowledge exchange, improved administrative efficiency, and increased semantic interoperability. The topics covered include ontology editors, real-world applications, and ontology development. Web miners employ techniques such as content mining and web use mining to generate ontologies, categorize individuals and sites, and retrieve conceptual linkages [9].

This study investigates the influence of ChatGPT on education through the utilization of web mining and natural language processing methodologies. Research reveals that ChatGPT significantly improves students' writing skills and fosters engaging and interactive learning settings. Nevertheless, it also gives rise to ethical considerations around plagiarism and academic dishonesty. The study indicates the necessity of establishing rules and policies for the implementation of AI tools in education. The results have both theoretical and practical consequences for incorporating ChatGPT into schooling [10].

Table 1: Web Usage Mining Methodology, Techniques and Challenges

Phase	**Steps**	**Techniques**	**Challenges**
Data Collection	Gathering web server logs	Log analysis tools	Data volume, and privacy concerns
Data Pre-processing	Data cleaning, sessionization, user identification	Data cleaning algorithms, sessionization techniques, user tracking methods	Data quality, noise, incomplete data
Pattern Discovery	Association rule mining, clustering, classification, sequential pattern mining	Data mining algorithms, statistical methods	Pattern interpretation, scalability
Pattern Analysis	Visualization, interpretation, application	Data visualization tools, statistical analysis, business knowledge	Actionable insights, resource allocation

This text presents a comprehensive examination of web usage mining, a discipline that entails the analysis of user behavior data obtained from websites. The text presents many strategies for discovering patterns, including association rule mining, sequential pattern mining, clustering, and classification. It also discusses ways for analyzing patterns, including as visualization and statistical analysis.

Table 2: Data Mining and Analysis Tools

Tools	**Description**
Statistical Software (e.g., R, SPSS, Python with libraries like Scikit-learn, Statsmodels)	For statistical analysis, hypothesis testing, and modeling
Data Mining Algorithms (e.g., Association Rule Mining, Clustering, Classification)	For discovering patterns and relationships in data

Machine Learning Libraries (e.g., Scikit-learn, TensorFlow, PyTorch)	For building predictive models
Visualization Tools (e.g., Tableau, Power BI, Matplotlib, Seaborn)	For creating interactive visualizations to explore data

Exploring user actions has never been easier than with Web Usage Mining, which has evolved from basic statistical analysis to state-of-the-art machine learning techniques. As data mining and ML continue to develop, the methods are also evolving at a rapid pace.

Table 3: Data Collection and Preprocessing

Tool	**Description**
Apache Web Server	A standard web server that generates log files
Log Analysis Tools (e.g., Splunk, ELK Stack)	For parsing, filtering, and aggregating log data
Data Cleaning and Preparation Tools (e.g., Python, R, Pandas, NumPy)	For handling missing values, outliers, and data transformation

Web Usage Mining is a powerful tool for extracting useful information from server logs for websites. Businesses can utilize Web Usage Mining to improve website design, personalize user experiences, and make data-driven decisions by analyzing patterns of user activity.

3.1 Research Gap

Although web use mining has made substantial progress in comprehending user behavior and enhancing web experiences, some crucial areas of research still need to be addressed. These tasks involve modifying techniques to accommodate changing data formats, facilitating analysis in real-time, and resolving difficulties in tracking across different devices.

Tracking user behavior accurately poses a difficulty when people engage with websites on various platforms such as smartphones, tablets, and desktops. There is much potential in analyzing user intent and sentiment from web interactions, even though most web usage mining focuses on user activity. A more complete picture may emerge if sentiment analysis methods were integrated with consumption statistics.

There has been phenomenal development in web mining as a whole. Web content mining, structure mining, and use mining are the three primary components that make up the field. The process of mining web usage logs for patterns in user access is known as web usage mining. The goal of web structure mining is to extract useful information from the structure of linked web pages. The process of gleaning useful information from online page content is known as web content mining.

4. CONCLUSION

Web mining is the process of extracting useful information from data stored on the internet, including web documents, links between websites, and usage logs. This data is obtained by applying data mining techniques.

Web mining is a field that specifically deals with studying the behavior of users on the internet, known as Web usage mining. Despite the novelty of the domain, we conducted a comprehensive review of the endeavors in this field, but with limited length. Web Usage Mining dynamic field with the potential to reshape the digital landscape. By harnessing its power responsibly and effectively, organizations can gain a competitive edge, enhance customer experiences, and drive meaningful business outcomes. Overall, Web Usage Mining has proven it can propel innovation and boost company performance. The importance of WUM in extracting value from web data will rise in tandem with the rate of technological development and the amount of data being generated.

In the future, the integration of real-time data analytics, the incorporation of more sophisticated machine learning algorithms, and the exploration of new data sources like social media are promising areas for future research. Additionally, ethical considerations and user consent will become increasingly important in the era of big data.

Author Contributions

The submitted version of the article was approved by all authors and all authors contributed to it.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCE

[1] Kumar, Avinash, et al. "Web Mining and Web Usage Mining for Various Human-Driven Applications." Advanced Practical Approaches to Web Mining Techniques and Application, edited by Ahmed J. Obaid, et al., IGI Global, 2022, pp. 138-162. https://doi.org/10.4018/978-1-7998-9426-1.ch007

[2] K. Griazev and S. Ramanauskaitė, "Web mining taxonomy," *2018 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania, 2018, pp. 1-4, doi: 10.1109/eStream.2018.8394124.

[3] R. K. Shukla, P. Sharma, N. Samaiya and M. Kherajani, "WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining," *2nd International Conference on Data, Engineering, and Applications (IDEA)*, Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170690.

[4] N. N. Akram and V. Ilango, "Intelligent Web Mining Techniques using Semantic Web," *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, Trichy, India, 2022, pp. 1-7, doi: 10.1109/ICEEICT53079.2022.9768546.

[5] L. Zhao and W. Jin, "Application of Web Data Mining and Information Combination Technology in E-commerce," *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2022, pp. 1636-1639, doi: 10.1109/ICEARS53579.2022.9752152.

[6] S. Rathi, Y. Deshpande, S. Nagaral, A. Narkhede, R. Sajwani and V. Takalikar, "Analysis of User's Learning Styles and Academic Emotions through Web Usage Mining," *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, 2021, pp. 159-164, doi: 10.1109/ESCI50559.2021.9397037.

[7] Castellano, G., Fanelli, A.M., Torsello, M.A. (2013). Web Usage Mining: Discovering Usage Patterns for Web Applications. In: Velásquez, J., Palade, V., Jain, L. (eds) Advanced Techniques in Web Intelligence-2. Studies in Computational Intelligence, vol 452. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33326-2_4

[8] B. Ravinder, S. K. Seeni, V. S. Prabhu, P. Asha, S. P. Maniraj and C. Srinivasan, "Web Data Mining with Organized Contents Using Naive Bayes Algorithm," *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Indore, India, 2024, pp. 1-6, doi: 10.1109/IC457434.2024.10486403.

[9] T. Zaidi, A. Kumar and S. Pundeer, "Shifting from Syntactic to Semantics for Knowledge Exemplification Using Semantic Web (SW) Mining Techniques," *2024 International Conference on Communication, Computer*

Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 2024, pp. 167-171, doi: 10.1109/IC3SE62002.2024.10593046.

[10] Abderahman Rejeb, Karim Rejeb, Andrea Appolloni, Horst Treiblmaier, Mohammad Iranmanesh, Exploring the impact of ChatGPT on education: A web mining and machine learning approach, , The International Journal of Management Education, Volume 22, Issue 1, 2024, 100932, ISSN 1472-8117, https://doi.org/10.1016/j.ijme.2024.100932.

[11] Horvat, M.; Krtalić, A.; Akagić, A.; Mekterović, I. Ontology-Based Data Observatory for Formal Knowledge Representation of UXO Using Advanced Semantic Web Technologies. Electronics 2024, 13, 814. https://doi.org/10.3390/electronics13050814

[12] Alqahtani, S.I.; Yafooz, W.M.S.; Alsaeedi, A.; Syed, L.; Alluhaibi, R. Children's Safety on YouTube: A Systematic Review. Appl. Sci. 2023, 13, 4044. https://doi.org/10.3390/app13064044

[13] Lee, J.W.; Han, D.H. Data Analysis of Psychological Approaches to Soccer Research: Using LDA Topic Modeling. Behav. Sci. 2023, 13, 787. https://doi.org/10.3390/bs13100787

[14] Oh, M.; Ahn, C.; Nam, H.; Choi, S. New Trends in Smart Cities: The Evolutionary Directions Using Topic Modeling and Network Analysis. Systems 2023, 11, 410. https://doi.org/10.3390/systems11080410

[15] Bottino, L.; Settino, M.; Promenzio, L.; Cannataro, M. Scoliosis Management through Apps and Software Tools. Int. J. Environ. Res. Public Health 2023, 20, 5520. https://doi.org/10.3390/ijerph20085520

[16] Génio, J.; Trifan, A.; Neves, A.J.R. Knowledge Maps as Support Tool for Managing Scientific Competences: A Case Study at a Portuguese Research Institute. Publications 2023, 11, 19. https://doi.org/10.3390/publications11010019