Exploring the landscape of Web Data Mining: An In-Depth Research Analysis

ABSTRACT

The exponential growth of Web services and Web-based applications has led to an enormous volume of data, providing a rich source for mining valuable insights. Web mining differs from traditional data mining due to the unique nature of the data it handles. Web data exists in diverse forms, including web server logs, news pages, and hyperlinks. As the usage of the internet continues to surge, web mining has become essential to extract meaningful information and patterns from these varied data sources. Traditional data mining methods may not be directly applicable to web data due to its unstructured and heterogeneous nature. Web server logs contain valuable information about user interactions, click-streams, and user preferences, which can be mined to understand user behavior and improve website performance. News pages and other forms of web content are valuable sources for sentiment analysis, topic modeling, and information retrieval, helping businesses and researchers gain insights into public opinions and trends. Additionally, web structure mining deals with the analysis of hyperlinks, enabling the discovery of relationships between web pages and identifying authoritative sources. The continuous growth of web-based data necessitates the use of specialized methods in web mining to effectively extract knowledge and valuable patterns. Researchers and practitioners in this field are constantly exploring innovative techniques to make sense of the vast amount of data available on the World Wide Web. The paper provides web mining techniques on web data and presenting the latest advancements, researchers and practitioners can gain insights into the state of the field and identify potential areas for further exploration. This paper also reports the comparisons and summary of various methods of web data mining with applications, which gives the overview of development in research and some important research issues.

Keywords: Information Retrieval, Semantic Web, Text Mining, Web Crawling, Web Mining, Web Content Mining, Web Data Mining, Web Structure Mining, Web Usage Mining.

INTRODUCTION

"The rapid growth of data in computer files and databases has led to increasing demands for more sophisticated information. Simple queries and structured query languages are no longer adequate to support these needs. Data mining steps in to find hidden information in databases and provide insights into complex relationships and patterns. Data mining is also referred to as exploratory data analysis, data-driven discovery, and deductive learning. In the data mining community, three types of mining are recognized: data mining, web mining, and text mining. Each type presents unique challenges and requires creative application of mining techniques" [1]. "Web mining, in particular, deals with semi-structured and unstructured data found on the web, which can be overwhelming due to its vast and heterogeneous nature. Web data mining involves the use of data mining techniques to extract valuable knowledge from web data, considering the hyperlink structure of the web or web log data, or both. Despite the absence of an agreed-upon definition, web data mining remains a crucial area of research for scholars in data mining and data management. Web Data Mining research is an ever-evolving field due to the dynamic nature of the web and the continuous growth of online data. Researchers strive to develop innovative and efficient methods to leverage the wealth of information available on the web for various applications and industries" [3]. "Interdisciplinary collaborations with fields like natural language processing, machine learning, and information retrieval continue to enrich the progress in Web Data Mining research. Data mining addresses the need for sophisticated information from growing databases, and web data mining tackles the challenges of handling heterogeneous and unstructured data found on the web. Researchers continue to explore innovative techniques to extract meaningful insights from this vast pool of web data. Data on the web is rapidly increasing day by day and Web data is huge, diverse and dynamic so information users could encounter the following problems while interacting with the web" [5].

1. Finding Relevant Information- People either browse or use the search service when they want to find specific information on the web. However today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.
2. Creating new knowledge out of the information available on the web-This problem is basically sub problem of the above problem. Above problem is query triggered process(retrieval oriented) but this problem is data triggered process that presumes that we already have collection of web data and we want to extract potentially use full knowledge out of it.
3. Personalization of information- When people interact with the web they differ in the contents and presentations they prefer.
4. Learning about Consumers or individual users-This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to website design and management and marketing etc.

WEB MINING OVERVIEW

In 1996, Etzioni [6] introduced the term "web mining" and hypothesized that the information on the web is sufficiently structured. He outlined the subtasks of web mining, describing it as the utilization of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Following in the footsteps of Kosala and Blockeel [5], Qingyu Zhang and Richard S. Segall [2] proposed breaking down web mining into several subtasks:

1. **Resource Discovery:** This involves locating unfamiliar documents and services on the World Wide Web.
2. **Information Selection and Pre-Processing:** This step focuses on automatically extracting and pre-processing specific information from newly discovered web resources.

3. **Generalization:** Uncovering general patterns at individual web sites and across multiple sites.
4. **Analysis:** Validating and interpreting the patterns mined from web data.
5. **Visualization:** Presenting the results of interactive analysis in a visual and easily understandable manner.

These subtasks form the core components of web mining, facilitating the efficient extraction and analysis of valuable information from the vast and diverse world of the internet. Researchers and practitioners continue to explore and develop techniques to enhance web mining and leverage its potential for various applications across domains.

According Kosala and Blockeel [5], web mining categorizes in to three categories depending on which kind of data to be mined that is mining for information or mining the web link structure or mining for user navigation patterns. These web mining texonomies are: (1) web content mining (2) web structure mining and (3) web usage mining shown below in figure 1.

Web content mining involves extracting valuable information from various types of web content, such as text, images, audio and video. On the other hand web structure mining focuses on leveraging hyperlinks to assess web page quality and make use of the collective judgment of the web pages. Web structure mining discovers underlying link patterns on the web, utilizing the topology of hyperlinks. Markov chain models aid in categorizing web pages and generating insights like website similarity and relationships. Web usage mining focuses on user navigation patterns, analyzing data from web server access logs, proxy server logs, user profiles and other interactions for user behavior discovery [9].

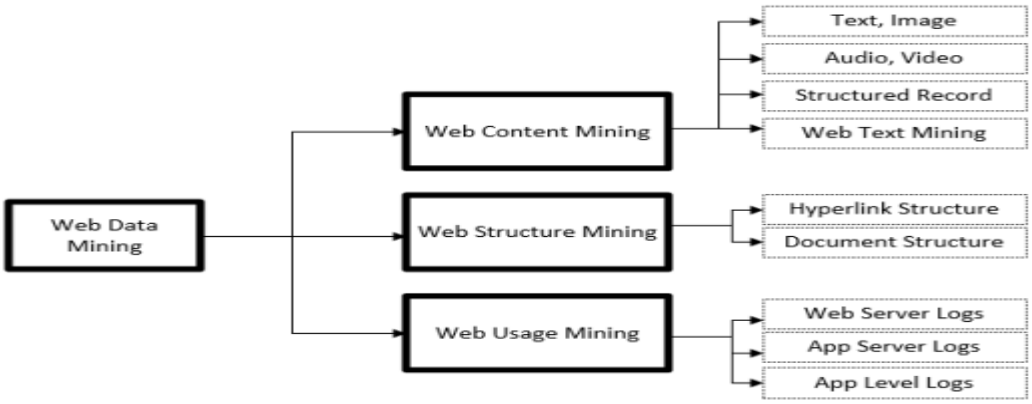

Fig.1. Web mining taxonomy

The incorporating data mining into web-page ranking benefits web search engines by identifying high quality web pages and improving web click-stream analysis. Data semantics play a crucial role in enhancing keyword based searches and addressing the challenges posed by unindexed documents. Leveraging data mining for web intelligence involves tasks such as mining web search engine data, analyzing web link structures, automatic classification of web documents, and exploring web page semantic structures and contents [8]. Additionally, web dynamics encompass studying how the web changes concerning its contents, structure and access patterns.

LITERATURE REVIEW

Web Content Mining

Web content mining focuses on discovering valuable information or knowledge from the content of web pages. Margaret H. Dunham [1] stated that It can be considered an extension of the work performed by basic search engines. Web content mining includes analyzing the content of web resources, such as text, images, sound, and video, to widen access to multimedia data. The primary web resources mined in web content mining are individual web pages. Information Retrieval is a research area that offers popular and effective statistical methods for web content mining. These methods can be used for document grouping, categorization, analysis, and retrieval. Web images lack proper manual annotations with semantic descriptors, making keyword-based image retrieval tools limited in their application. Traditional text-based methods also struggle to handle the vast number of images available. Content-Based Image Retrieval

(CBIR) was developed to address these challenges. CBIR automates the process of indexing or annotating images in image databases using visual contents such as color, shape, texture, and spatial layout. These visual features are represented as multidimensional feature vectors and indexed in a database [4]. “When an image is used as a query, its feature vectors are extracted, and similar images are retrieved from the database through comparison. The indexing scheme ensures an efficient image retrieval process. A significant challenge in content based image retrieval (CBIR) systems is the gap between low-level image features, such as color histograms used for indexing, and high – level semantic content, which is subjective to human perception. To address this gap, Zhang propose using relevance feedback techniques to refine queries or similarity measures in the image search process, bridging the divide between image features and semantic understanding in CBIR systems or search engines” [21].

In content-based image retrieval (CBIR), relevance feedback is utilized to enhance retrieval performance based on user-provided feedback examples. Zhang et al [21]. introduced a framework that integrates low-level features and keyword explanations in the retrieval and feedback processes, aiming to improve the effectiveness of the information system. Initially, relevance feedback approaches focused on low-level features, essentially replacing keywords with features for document retrieval. However, relying solely on low-level features may not efficiently capture users' feedback and intentions. To interact with the image retrieval system, users have two options. First, they can input a list of keywords representing the semantic contents of their preferred images. Second, they can provide a set of example images, and the system retrieves analogous images. Combining these two approaches allows for improved retrieval accuracy and enhances the system's user-friendliness.
The CBIR framework with integrated relevance feedback and query expansion comprises the following key steps:

1. **Semantic Network:** This component links images to semantic annotations stored in a database. The system incorporates a similarity measure that combines both semantic and low-level image features. Additionally, a machine learning algorithm is employed to iteratively update the semantic network, leading to continuous improvement in the system's performance over time.
2. **Query Support:** The system supports both query by keyword and query by image example using the semantic network and low-level feature indexing.

Figure 2 illustrates the integrated relevance feedback and query expansion framework, while Figure 3 displays the semantic network. These components work in tandem to enable efficient and effective content-based image retrieval, leveraging both semantic understanding and low-level image features. To further enhance the retrieval performance of the proposed framework, a cross-modality query expansion method is implemented. When a query is submitted using keywords, the retrieved images from the keyword search are considered as positive examples [15]. The query is then expanded using the features of these images. This process involves searching the semantic network for the keywords and incorporating the visual features of the images containing these keywords (referred to as training images) into the expanded query. The quality of the system's retrieval performance improves with a higher number of correctly annotated images. However, manual image labeling is laborious and costly, making it an impractical solution. To address this challenge, H. J. Zhang introduces a probabilistic progressive keyword propagation scheme in the framework [21]. This scheme automatically annotates images in the databases during the relevance feedback process using a small percentage of annotated images. This approach significantly reduces the burden of manual labeling and enables the system to iteratively improve its performance through automated image annotation.

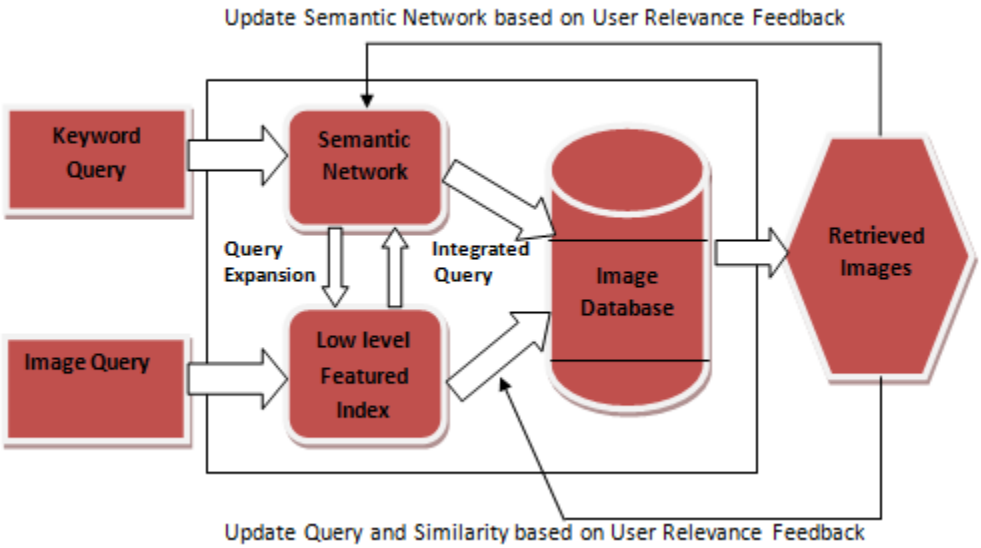

Fig.2. Framework of integrated relevance feedback and query expansions

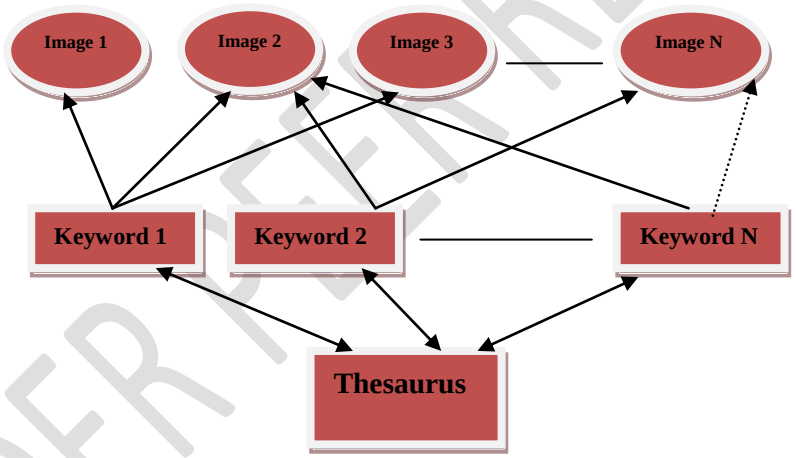

Fig.3. Semantic network

The researchers developed iFind, an Image Search Engine, which outperforms traditional approaches. iFind supports various search options, including Keyword-based search, Query by example, Relevance feedback, and log mining. Chen et al. [22] proposed a novel representation technique called Semantic Virtual Document (SVD), which utilizes Web structure and summarization techniques to better represent knowledge in actual Web Documents. SVD, when combined with a suitable clustering algorithm, enables automatic content-based categorization of similar web documents. This technique facilitates automatic content-based classification of web documents and offers a tree-like graphical user interface for browsing, enhancing the relevance judgment process for internet users. In their work, they also introduce a cluster-biased automatic query expansion technique to accurately interpret short queries. The researchers present “a prototype called iSEARCH for web content mining, providing an intelligent search and review of cluster hierarchy. Most existing web mining algorithms have primarily focused on identifying frequent patterns, overlooking less frequent ones that may contain outlying data, such as noise, irrelevant, and redundant information” [13]. “The retrieval of relevant content from the web is a common task, necessitating the development of user-friendly and automated tools to provide relevant information efficiently. Improving web content mining results involves detecting and eliminating redundant links, which

poses a significant challenge in current web mining research" [18]. This summarizes various methods of web content mining along with their applications in table 1.

Table 1 Summarizes various methods of web content mining along with their applications

Author	Method	Application
Dr. Fuhui Long	Visual content description	Content based image retrieval
H. Zhang	Relevance feedback algorithm	Content based image retrieval (iFind)
Chen	Web structure together with summarization techniques	Semantic virtual document (iSearch)
Ricardo Campos	Graph-based overlapping, Clustering algorithm	Meta-search engine called WISE
Mehdi Hosseini	Query-URL co-clustering	Categorize queries and URLs related to special website
Hui Zhang	SPARSE technique	Localized CBIR system,
G. Poonkuzhali	Signed approach and full word matching	Retrieval of documents takes less time and less space

Web Structure Mining

"Web structure mining is concerned with discovering and modeling the link structure of the web. While web information retrieval tools typically rely on the text available on web pages, they often overlook valuable information contained within web links. Web structure mining aims to generate structural summaries about web sites and web pages, with a primary focus on link information" [12]. By utilizing web structure mining techniques, it becomes possible to discover various aspects, such as visible web documents, luminous web documents (those with a significant number of outgoing links), and luminous paths (sets of interlinked nodes). Many search engines fail to uncover such valuable insights, making web structure mining a powerful tool for revealing hidden relationships and patterns in the link structure of the web. Kleinberg introduced the HITS algorithm (Hypertext Induced Topic Search), an original approach for hyperlink analysis. He proposed the concepts of "hubs" (pages that refer to many other pages) and "authorities" (pages that are referred to by many other pages). With these concepts, Kleinberg developed a set of algorithmic tools to extract valuable information from the link structures of hypertext environments like the World Wide Web [3]. The main objective of the HITS algorithm was to refine broad search topics by identifying authoritative information sources related to those topics. Through experiments conducted on the World Wide Web, Kleinberg demonstrated the effectiveness of the HITS algorithm in various contexts. The algorithm's ability to discover hubs and authorities allowed for more accurate and targeted information retrieval, making it a significant advancement in hyperlink analysis for information retrieval on the web [6]. Furnkranz [19] described the web as a directed graph, with documents as nodes and hyperlinks as edges, which can be exploited for better ranking and classification in search engines. To assist users in searching and organizing information on the web, Smith and Ng suggested using a self-organizing map (SOM) to mine web data and developed LOGSOM, a system that organizes web pages into a two-dimensional map, providing a meaningful navigation tool for users. The design of web site portal pages by proposing a heuristic approach called Link Selector for hyperlink selection. They aimed to maximize the efficiency and effectiveness of portal pages in presenting information. Hay used the Sequence Alignment Method (SAM) to partition users into clusters based on web page request

sequences, demonstrating its effectiveness in identifying similar behavioral patterns among users. We designed a new bookmark structure to store web object references, allowing easy access from anywhere on the internet. They proposed a prototype with additional features to share bookmarks among groups of users. Song and Shepperd used vector analysis and fuzzy set theory to cluster users and URLs and identify frequent access paths for e-commerce websites. Nacim Fateh Chikhi, Bernard Rothenburger investigated the equivalence between HITS and Principal Component Analysis (PCA) for dimensionality reduction [20]. They compared various dimensionality reduction techniques (PCA, Non-negative Matrix Factorization, Independent Component Analysis, and Random Projection) for web structure mining and found Non-negative matrix factorization to be a promising approach for web structure analysis. Overall, these studies demonstrate the ongoing efforts to leverage graph structures and advanced algorithms for enhanced web content mining and retrieval tasks.

Table 2: Summarizes various methods of Web structure mining along with their applications

Author	Method	Application
Sanjay Kumar Madria	Warehouse of Web Data (WHOWEDA project)	To design the tools and techniques for web data mining
Kleinberg	HITS	Discovering authoritative sources in a hyperlinked environment
Johannes Furnkranz	Data mining and machine learning	Exploiting the graph structure of the Web
Smith and Ng	Clustering, self-organized map	Mapping user navigation patterns (LOGSOM)
Fang and Sheng	Heuristic approach	Hyperlink selection for portal page
Hay	Sequence Alignment Method (SAM)	Mining navigation Patterns
Guan and McMullen	Design bookmark structure	Bookmark
Song and Sheppard	Frequent access path identification algorithm, fuzzy set theory	Mining Web Browsing patterns for e-commerce
Nacim Fateh Chikhi	Various dimensionality reduction techniques (DRTs)	To extract the implicit structures hidden in the web hyperlink connectivity
Lefteris Moussiades	Graph clustering algorithm	Mining the community structure of a graph

Hyperlinks serve a dual purpose, not only aiding search ranking but also facilitating the discovery of Web communities. Web communities consist of web pages focused on specific topics or themes. Many community mining approaches operate under the assumption that community members have more hyperlinks within the community than outside it. As a result, graph clustering algorithms are often employed to mine the community structure of a graph, as they share the same assumption, considering clusters as subsets of vertices with higher internal link density than external link density. Various methods of web structure mining and their applications have been summarized in Table 2. “Web Structure Mining plays a crucial role, offering several benefits, including quick responses to web users and reducing HTTP transactions between users and servers” [14]. The survey on web structure mining aims to provide valuable information concerning knowledge issues on the web. It offers insights into discovering meaningful patterns and relationships within web data, contributing to a better understanding of web communities and content organization.

Web Usages Mining

Web usage mining is concerned with understanding user behavior while interacting with the web or a website. Its primary goal is to gather information that can assist in web site reorganization or adapt the

site to better suit user needs. This type of mining focuses on analyzing server logs to extract valuable insights into users' access patterns and preferences [10]. “By leveraging the information in log records, web sites can optimize themselves to meet users' requirements, enhance user experience, and achieve better economic benefits. Numerous web log analysis tools are available to mine data from log records, which contain essential information such as URLs, IP addresses, and timestamps. Analyzing and discovering log data enables organizations to identify potential customers, determine page popularity (based on the number of visits), and gain insights for reorganizing the web site to facilitate faster and easier customer access. Additionally, web log analysis aids in improving links and navigation, attracting more advertisement revenue through intelligent adverts, converting viewers into customers through better site architecture, and monitoring the overall efficiency of the web site. By harnessing web usage mining, businesses can make data-driven decisions and enhance their online presence to meet the evolving needs of their users effectively” [11].

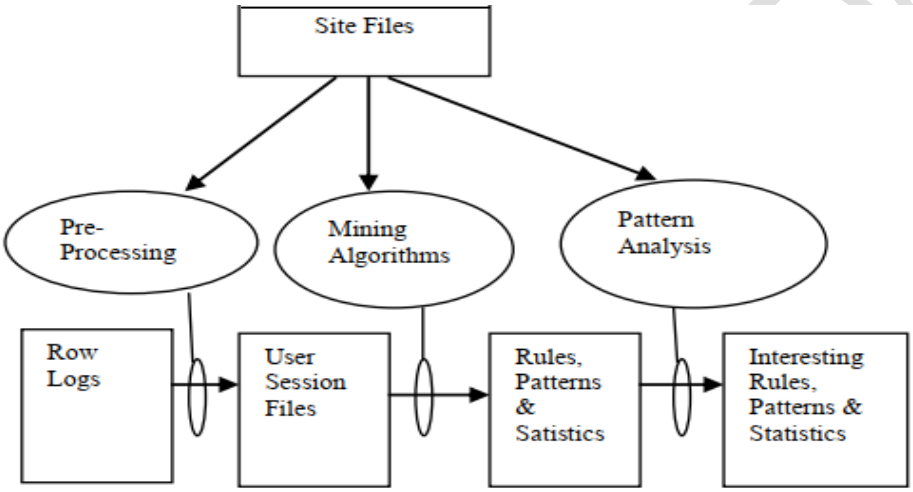

Fig.4. High level web usage mining process

Web mining relies on data collected from various sources, including Web servers, clients, proxy servers, and server databases. However, these sources often produce noisy data, which can significantly impact the accuracy of web mining results. To address this, data cleaning methods are crucial in the preprocessing stage of web mining. Jaideep Srivastava and R. Cooley categorized data preprocessing into subtasks, aiming to obtain clean data that can identify a user's browsing pattern, such as page views, sessions, and click streams. Click streams, in particular, are valuable for reconstructing user navigational patterns. Markov models have been widely used to model users' navigation behaviors on web sites. Jianhan Zhu, Jun Hong [17] introduced the Citation Cluster clustering algorithm to group conceptually related pages and create a conceptual hierarchy of the web site. This hierarchy is then integrated with Markov model-based link prediction to enhance users' navigation experience on the web site. In the past six years, various models for representing user navigation sessions have been proposed, such as Hyper Text Probabilistic Grammar (HPG), N-Gram Model, and Dynamic clustering-based Markov model. These models play a crucial role in understanding user behavior and improving web site navigation and personalization for users. Web usage mining (WUM) algorithms are designed to utilize Web log records to discover valuable knowledge that can support business applications and decision-making processes. However, the effectiveness of WUM in knowledge discovery depends on both the algorithm used and the quality of the data. In a study conducted by Yu-Hui Tao, Tzung-Pei Hong [18], a new data source called intentional browsing data (IBD) is explored to potentially enhance the effectiveness of WUM applications. IBD refers to a category of online browsing actions, such as "copy," "scroll," or "save as," which are not typically recorded in standard Web log files. The research aims to develop a better understanding of IBD and facilitate its seamless integration into WUM research and practical applications. By exploring this new data source, researchers aim to uncover valuable insights that were previously overlooked and improve

the overall performance of web usage mining algorithms in real-world scenarios. We have summarized various methods of web usage mining with application in Table 3.

Table 3: Summarizes various methods of Web usage mining along with their applications

Author	Method	Application
Jaideep Srivastava, R. Cooley	Statistical Analysis Association Rules	Personalization Site Modification
Jianhan Zhu	Clustering algorithm called Citation Cluster	Construct a conceptual hierarchy of the Web site
Borges and M. Levene	Dynamic clustering based method	Representing a collection of user web navigation sessions
TAN Xiaoqiu, YAO Min	Improved WAP tree	Sequential pattern mining
Yu-Hui Tao , Tzung-Pei Hong	Taxonomy of browsing data	Decision support
Mehdi Hosseini	Web based recommender systems	predict user‘s intention and their navigation behaviors
Mehrdad Jalali	Longest common subsequences algorithm	Predict user near future movement.
M. Jalali	WebPUM	Predict user near future Movement

CONCLUSION AND FUTURE DIRECTIONS

This paper presents about ongoing web mining research scenarios, mainly focusing on three types of web mining: web content mining, web usage mining, and web structure mining. The review encompasses an extensive literature survey, offering insights into the progress made in each type of mining over the years. Web data mining has emerged as a rapidly growing research area, and its significance is poised to increase further as web data and its usage continue to expand. The ever-increasing volume of web content, structure, and usage data creates new challenges and opportunities for researchers in this field. Web data is predominantly semi-structured to unstructured, which poses significant challenges due to the heterogeneity and lack of structure. Consequently, the automated discovery of targeted or unexpected knowledge information remains a daunting task, presenting several complex research problems. While a substantial portion of knowledge is represented in HTML web documents, it is essential to consider the various other file formats accessible on the internet. Additionally, if web documents and corresponding backlink documents consist mainly of multimedia information such as graphics and audio, traditional techniques like Singular Value Decomposition (SVD) might not be optimally effective in revealing textual information. Therefore, there is a need to explore new techniques that can incorporate these diverse file formats and multimedia information for better knowledge representation. Despite significant progress, web data mining is still in its early stages, and a plethora of ongoing research indicates the field's potential for further advancements. As the importance of web data continues to grow, researchers must continue their efforts to tackle the existing challenges and explore novel solutions to drive the field

forward. Furthermore, the paper also reports on comparisons and summaries of various methods used in web data mining, along with their applications. By offering an overview of research developments, it sheds light on the progression of the field and identifies critical research issues that warrant attention. In conclusion, this paper serves as a valuable resource for researchers and practitioners in web data mining, providing a comprehensive understanding of past achievements and paving the way for innovative research in the future.

REFERENCES

1. Margaret H. Dunham, ―Data Mining Introductory & Advanced Topics‖, Pearson Education.
2. Qingyu Zhang and Richard s. Segall, "Web mining: a survey of current research, Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720.
3. Mustafa Ali Bamboat, Ghulam Sarfaraz Khan, Naadiya Mirbahar, Sheeba Memon, "Web Content Mining Techniques for Structured Data: A Review" (SJHSE) Sindh Journal of Headways in Software Engineering, Volume 01, Issue 01 (2022)
4. Andemariam Mebrahtu, and Balu Srinivasulu, "Web Content Mining Techniques and Tools," 2017 – IJSCMC; URL: https://www.ijcsmc.com/docs/papers/April2017/V6I4201725.pdf
5. A. Richlin Selina Jebakumari , Dr. Nancy Jasmine Goldena, "A Survey on Web Content Mining Methods and Applications for Perfect Catch Responses" International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 01 | Jan 2019 e-ISSN: 2395-0056 p-ISSN: 2395-0072, PP 407-412.
6. Kosala and Blockeel, "Web mining research: A survey",SIGKDD:SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000.
7. O.etzioni. The world wield web: Quagmire or Gold Mining. Communicate of the ACM, (39)11:65-68, 1996.
8. Kumar, S., & Kumar, R. (2021)." A Study on Different Aspects of Web Mining and Research Issues. In IOP Conference Series: Materials Science and Engineering" (Vol. 1022, Issue 1, p. 012018). IOP Publishing. https://doi.org/10.1088/1757- 899x/1022/1/012018.
9. Sharma, P. S., Yadav, D., & Thakur, R. N. (2022). "Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey". In M. P. Kumar Reddy (Ed.), Mobile Information Systems (Vol. 2022, pp. 1–19). Hindawi Limited. https://doi.org/10.1155/2022/7519573.
10. T. Sunil kumar1, Dr. K. Suvarchala, A Study: Web Data Mining Challenges and Application for Information Extraction, IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727Volume 7, Issue 3 (Nov. - Dec. 2012), PP 24-29.
11. J. Just, A Short Survey of Web Data Mining WDS'13 Proceedings of Contributed Papers, Part I, 59–62, 2013. ISBN 978-80-7378-250-4, MATFYZPRESS.
12. Ramakrishna, Gowdar et al "Web Mining: Key Accomplishments, Applications and Future Directions", in the International Conference on Data Storage and Data Engineering 2010.
13. Kavita, Mahani P, Ruhil N. Web data mining: a perspective of research issues and challenges. In international conference on computing for sustainable global development 2016 (pp. 3235-8). IEEE.
14. Singh B, Singh HK. Web data mining research: a survey. In international conference on computational intelligence and computing research 2010 (pp. 1-10). IEEE.
15. Anil B. Pawar, Madhuri A. Jawale, Chaitanya P. Kale, "A Powerful Techniques and Applications of Web Mining" in Intelligent Systems and Computer Technology, (pp. 269-276) doi:10.3233/APC200153.

16. Wang Bo and Xu Jing, "Research on Web Data Mining Hadoop Simulation Platform Based on Cloud Computing", Electronic Design Engineering, vol. 26, no. 2, pp. 22-25, 2018.
17. Jianhan Zhu, Jun Hong et al, ”Using Markov Models for Web Site Link Prediction” College Park, Maryland, USA ACM June 11-15, 2002.
18. Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su, ”Web usage mining with intentional browsing data” in international journal of Expert Systems with Applications 34 (2007), 1893–1904.
19. J. Furnkranz, Web structure mining — Exploiting the graph structure of the worldwide web, OGAI-J. 21(2)(2002) 17–26.
20. Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles, “A Comparison of Dimensionality Reduction Techniques for Web Structure Mining”, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,P.116-119, 2007.
21. H. Zhang, Z. Chen, M. Li and Z. Su, Relevance feedback and learning in content-based image search, World Wide Web 6(2) (2003) 131–155.
22. L. Chen, W. Lian and W. Chue, Using web structure and summarization techniques for web content mining, Inform. Process. Management: Int. J. 41(5) (2005) 1225–1242.