Hindawi
Mobile Information Systems
Volume 2022, Article ID 7519573, 19 pages
https://doi.org/10.1155/2022/7519573

Review Article

Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey

Prem Sagar Sharma,[1] **Divakar Yadav**,[2] **and R. N. Thakur**[3]

[1]*Uttarakhand Technical University, Dehradun, Uttarakhand, India*
[2]*National Institute of Technology, Hamirpur, Himachal Pradesh, India*
[3]*LBEF Campus, Kathmandu, Nepal*

Correspondence should be addressed to R. N. Thakur; rn.thakur@lbef.edu.np

Received 17 March 2022; Revised 8 April 2022; Accepted 4 May 2022; Published 31 May 2022

Academic Editor: M. Praveen Kumar Reddy

Due to the exponential growth of Internet users and traffic, information seekers depend highly on search engines to extract relevant information. Due to the accessibility of a large amount of textual, audio, video etc., contents, the responsibility of search engines has increased. The search engine provides relevant information to Internet users concerning their query, based on content, link structure, etc. However, it does not guarantee the correctness of the information. The performance of a search engine is highly dependent upon the ranking module. The performance of the ranking module is dependent upon the link structure of web pages, which analyze through Web structure mining (WSM) and their content, which analyzes through Web content mining (WCM). Web mining plays a vital role in computing the rank of web pages. This article presents web mining types, techniques, tools, algorithms, and their challenges. Further, it provides a critical comprehensive survey for the researchers by presenting different features of web pages, which are essential to check their quality. In this work, authors presented different approaches/techniques, algorithms and evaluation approaches in previous researches and identified some critical issues in page ranking and web mining, which provide future directions for the researchers working in the area.

1. Introduction

The size of web documents over the World Wide Web (WWW) has exponentially increased due to increasing the dependency of users over the Internet. An automatic system is required to fetch reliable information from such a huge collection of web documents because this task is challenging to analyze manually. Search engine [1–3] is an information retrieval tool for the Web like Google, Yahoo, Bing. The summary of various search engines is shown in Table 1. Still, these search systems can sometimes not guarantee reliable and accurate information, but still, these systems provide better results than performing the task manually by experts. These tools often do not provide precise information because the IR system [6] returns information to Internet users based on specific retrieval criteria. For instance, it fetches web documents based on the subject/title as given. To fetch huge web documents related to a specific domain is very easy and common. Therefore, search engines provide a ranking system to find reliable web documents for user/client queries. Generally, a ranking mechanism creates the rank of web pages based on either keywords/reliability or links/popularity.

Hyperlinked Structure [7] was developed in 1989 to share information among researchers in Switzerland. Later, it became a platform of WWW development guided by the WWW association at MIT (Massachusetts Institute of Technology) in Cambridge. The recent growth of WWW has changed the computer science & engineering and the people's lifestyles and economics of various countries.

Since its onset, the WWW has been increasing exponentially as shown in Figure 1(a) A 10–106 terabytes of traffic have increased in a month between 1995 and 2000. The total web traffic between 2005 and 2010 increased from 1 to 7 exabytes. Now in 2020, Internet traffic is increasing

Table 1: History of various search engines [2].

Search engine	Year	Methodology	Description
Gerard Salton [4]	1960s–1990s	—	Vector space model (VSM), inverse document frequency (IDF), term frequency (TF), term Discrimination values (TDV), and feedback mechanisms
Archie and Veronica [4]	1991–1992	—	"Archie" work for FTP sites & "Veronica" work for Gopherspace. Gopherspace describes the aggregate of whole the information (such as document file, papers, abstracts, and other types of files) on the various Gopher servers in the world
The first web directory [4]	1993–1994	—	ALIweb (archie like indexing) was created in October 1993 due to an automated indexing problem. The first time it was creating a directory for the web. It (directory) stores URLs and their description
Search directories [4]	1994–1995	It use large database which is created automatically	It was the first browser-based web directory. It is also doing help to the user to coordinate directories. "Yahoo" becomes more popular, which provide an interface to make easy interaction for the user
Yahoo [5]	1995	CORE (content optimization and relevance engine)	Yahoo! search engine was developed in 1995. It is written in PHP. Originally, crawling and data storage was not done by Yahoo! It was the first popular web search engine
Meta-engines [4]	1995	It uses fusion to filter data The two main fusion methods: 1-collection fusion, 2-data fusion	Meta engines play important role in search engines. There is nothing new but they work together to collect results from various search engines. it was introduced in 1995 in Washington
Ask [4]	1996	It works on the basis of asking question instead of one or more words	Ask.com also called Ask Jeeves is focused on question answering, mainly for e-business developed in 1996 by Garrett Gruener and David Warthen in Berkeley, California. The main task of this search engine was to rank links based on popularity
Google [5]	1997	PageRank	Larry Page and Sergey Brin developed the Google engine in 1996. Nowadays, Google is one of the most reliable search engines. It works based on web structure mining
Bing [4]	2009	Bing's ranking algorithm utilizes machine learning/AI	MSN web portal developed in 2005 for in house search, but renamed in 2009 by Bing web search engine owned & launched by Microsoft. It is also called Microsoft Bing. Initially it work for window live search and was later also used in live search. It was developed in ASP.NET

approximate 5.3 exabyte per day. According to Cisco, 82% of video-Internet traffic of all web traffic will be in 2021. In 2016, 73% of video traffic [8] of all Internets was present as shown in Figure 1(b). People view large amounts of video, but they also use high bandwidth to view good-quality videos.

All types of web content (like video, Netflix, webcam) generate demand. Now growing live videos is an integral part of the Internet. These video offerings from various sources like live Facebook, Twitter's broadcast, live YouTube, live sports is expected to increase approximately 13% of traffic as shown in Figure 1(c) of total video web traffic by 2021 [8]. WWW is an essential and widely used tool to provide reliable information to Internet users. It provides an essential and easy mechanism for information like static text, images, dynamic and interactive services such as audio/video conferences. It provides the facility to view various types of information, including magazines, library resources in different sectors, current and business news, etc. Now the web is an essential source of all kinds of information.

The information retrieval systems [9] were developed to store and search web pages in efficient manner because the size of WWW increased exponentially. Generally, the text documents are stored in text databases, and the IR system provides a framework to enable searching. The IR system generates a list of documents in response to a query. In general, these are listed in descending order by estimated relevancy. Because most users only glance at the first 10–50 items (the maximum criterion), the algorithms try to put the most relevant papers at the top.

However, searching for information on the web is difficult for an information seeker. Web-based information retrieval systems called search engines [10] have made things easy for information seekers but do not provide guarantees about the correctness of the information. Many times, the information is not precise. It is a program that searches for the documents for specified queries and returns the list of documents where the query keywords were found.

It is important to understand that the term "popularity" is normally the result of link analysis and not user feedback. A web search engine as shown in Figure 2, typically consists of a ranking system that measures the importance of Web Pages [11, 12]. Using the hybrid approach, one can fetch content-based information from web documents [13]. The traffic of search engines is affected [14] by the following factors: size of the web, loading speed [15], web security condition, SEO Crawling Factor (Title, heading, Meta Description of web page, Content, URL), User behavior [11, 16]. [17] presents a web

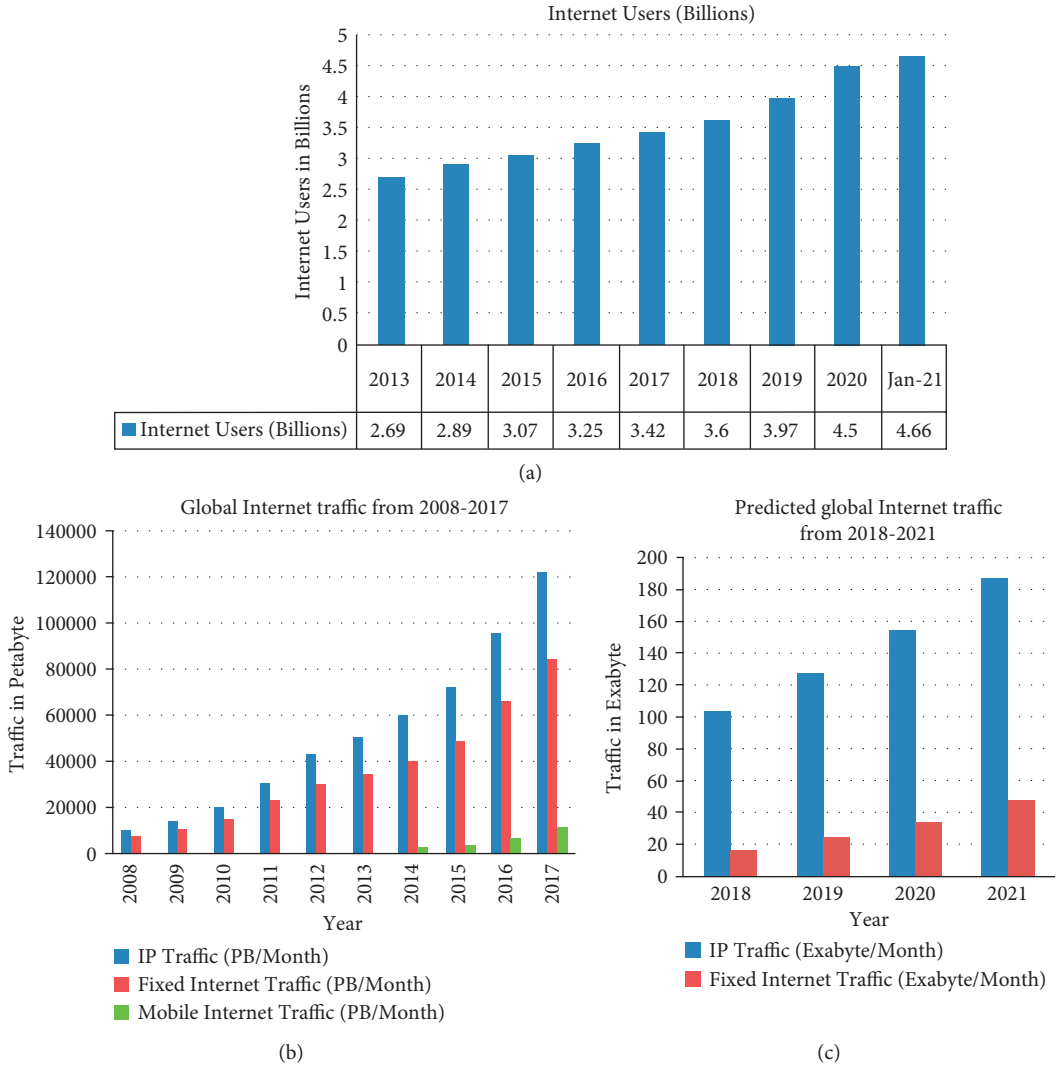

Figure 1: (a). The exponential growth in terms of actual total Internet users [1] (b). Global Internet traffic from 2008–17 (c) Predicted Global Internet traffic from 2018–21.

page rank mechanism that is query dependent. This approach was much better and effective, but it took more time to rank. In [18], the authors present a ranking mechanism based on link attributes, but it was not able to check the content quality of the web page. Some content-based ranking approaches are presented in [19–21]. The main issue in content mining is that it was increasingly perceived latency, addressed in [22] by an additional component, said the proxy server.

Search engines follow the following steps to process user queries:

(a) Take user query and, based on its keywords, make a precise query to process.

(b) Analysis and Fetch data from web repository corresponding user request.

(c) Ranked to all fetched web pages.

(d) Return the list of URLs array of ranked web pages for the user request.

(e) Get the updated user query of the user, if any?

1.1. Working Process of Search Engine

```
frontend_search_engine (UserQuery)
    {
(1) result_QP = Quesry_processor (UserQuery, Indexed_Web_Repository, Meta_data);/ *Fetch web record from web repository corresponding user query and store into "result_QP" */
(2) ranked_web_pages = Ranking_system (result_QP, Meta_data);/ * Ranking system process result of Query Processor "result_QP" to arrange all web pages into high rank to low rank */
    }
backend_search_engine (URL_List)
    {
(1) WebPageRepository = Crawler (URL_List);/ *Crawl the web pages by crawler with the help of robot.txt file, store into web page repository and
```

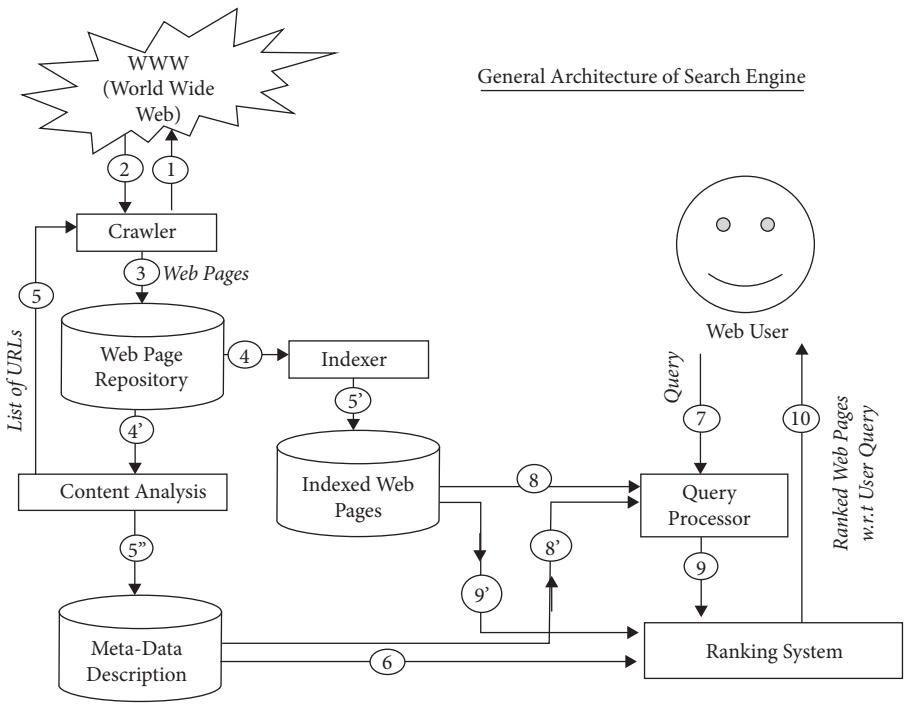

Figure 2: General architecture of search engine.

add new URLs into URL_List to crawl all these web pages also. */

(2) indexed_web_page_Repository = indexer (webPageRepository);/*The indexer analyses all extracted documents by extracting relevant terms for creating an index to search documents against user queries */

(3) new_list_of_URLs = contentAnalysis (webPageRepository); *Content Analysis compute the relevance of a web page on the basis of its contents with respect to user query */

(4) Meta_data = contentAnalysis (webPageRepository);

(5) Update_URL_List (URL_List, new_list_of_URLs); }

QueryProcessor (UserQuery, Indexed_Web_Repository, Meta_data)

{

(1) WebPageRepository = Crawler (URL_List);/ *Crawl the web pages by crawler with help of robot.txt file, store into web page repository and add new URLs into URL_List to crawl all these web pages also. */

(2) indexed_web_page_Repository = indexer (webPageRepository);/*The indexer analyzes all extracted documents by extracting relevant terms for creating an index to search documents against user queries */

(3) new_list_of_URLs = contentAnalysis (webPageRepository);/ *Content Analysis compute the relevance of a web page on the basis of its contents with respect to user query */

(4) Meta_data = contentAnalysis (webPageRepository);

(5) Update_URL_List (URL_List, new_list_of_URLs) }

2. Web Mining

Data mining is used to find out relevant patterns or knowledge from repositories (such as databases, texts, images), which should be valid, valuable, and understandable. Text mining has become popular and reliable by increasing the popularity of text documents. Web mining [23–25] is used to fetch useful/relevant information and use this information to generate knowledge and personalize the information and learn about users. The hyperlink structure of web pages, the content of web pages is used to collect the relevant information. Data mining techniques as shown in Figure 3 [25–28] are used to fetch and discover relevant information automatically from web pages and web services in web mining. Data mining services are discussed in [29] to extract something useful out of the Web. There are following steps are needed to perform for this purpose:

(i) *Resource finding:* Extract the useful data/resources from either web documents, which are available online, or offline mode.

(ii) *Information selection and preprocessing:* Apply the preprocessing (cleaning, normalization, feature extraction) on the specific information, which is automatically selected.

(iii) T*ransformation:* Preprocessed data are transformed into valuable information by removing stop words to obtain necessary phrases in training mass.

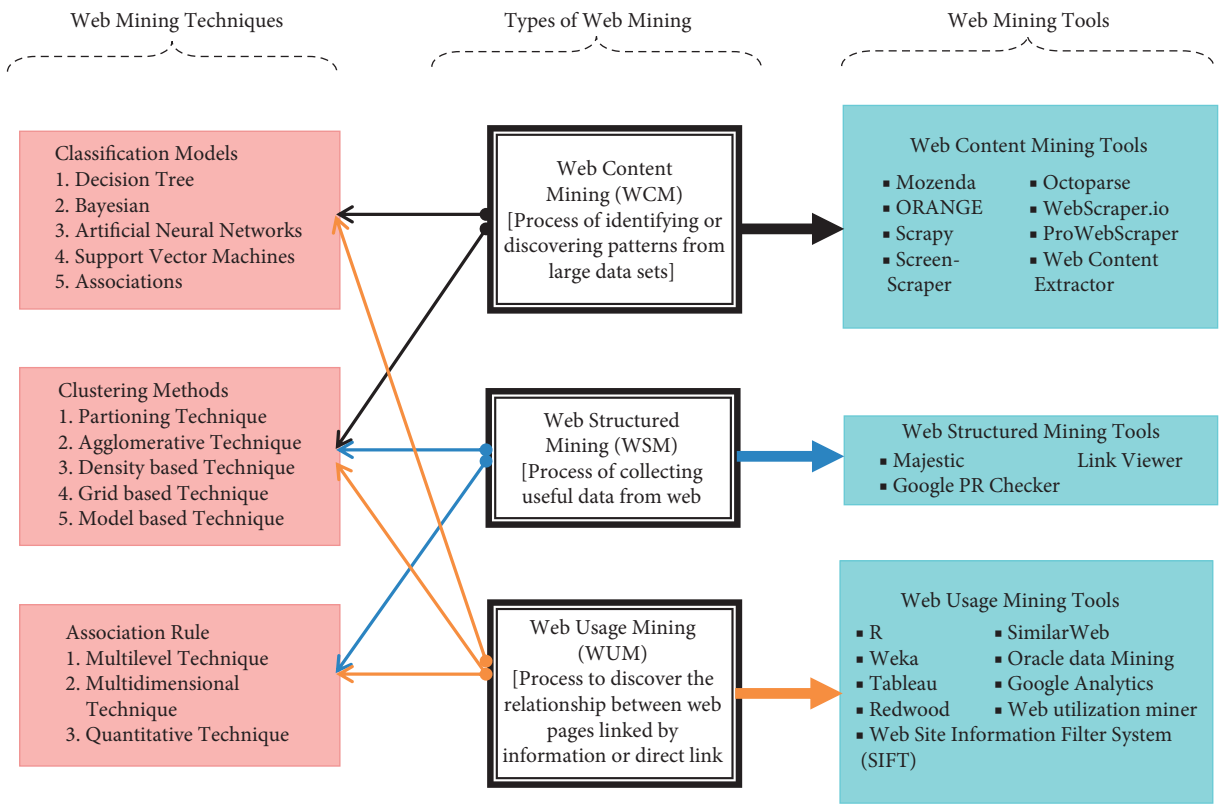

Figure 3: Summarization of web mining types, classification model, and tools.

(iv) *Generalization:* It is used to fetch patterns from a website or across various websites by applying machine learning (ML) and other data mining techniques.

(v) *Analysis:* This phase analyses mined patterns by validation and interpretation. Pattern mining plays an important role in this phase. In knowledge generation on the web, human being plays an important role.

There are three basic information such as the previous pattern, shared content's degree, and link structures in web mining discussed below:

2.1. Web Usage Mining (WUM). Web and application servers are the main sources to collect web log data. Log files generate over the web whenever an Internet user interacts with the web through search engines (shown in Figure 4).

The following techniques [3] are used in web usage mining:

2.1.1. Association Rules. By using association rule creation in the Web domain, pages that are frequently referred together can be combined into a single server session. Unordered correlation between objects observed in a repository of activities that can be discovered using association rule mining techniques. In web usage mining, the association rules apply to groups of pages that are accessed together and have a support value that is greater than a certain threshold. Support value is the percentage of activities for a specific pattern. The presence or absence of association rules can help Web designers rebuild their pages more effectively. Association rules can be used as a trigger for pre-fetching documents while loading a page from a distant site to reduce user perceived latency. Association rules in WUM provide the relationship between web pages that frequently appear next to one another in user sessions [6, 7].

Statement of association rules written as follows:

$$A => B, \quad (1)$$

where A, B are sets of items in a series of transactions.

For example, an association rule: Page A, Page B => Page C shows, if the user/client observe page A, and B then page C will be observed in the same meeting.

2.1.2. Classifications. Classification is used to map a data item into predefined classes.

In the Web domain, it is necessary to extract and select attributes that best characterize the properties of a specific class or category in order to create a profile of people belonging to that class or category. The web usage mining process understands the existing data and behavior of new instances. It identifies a particular class/category of a user. Classification techniques use Machine Learning (ML), Neural Network (NN) and statistical. Decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines, and other supervised inductive learning techniques can be used for classification as shown in Figure 5.

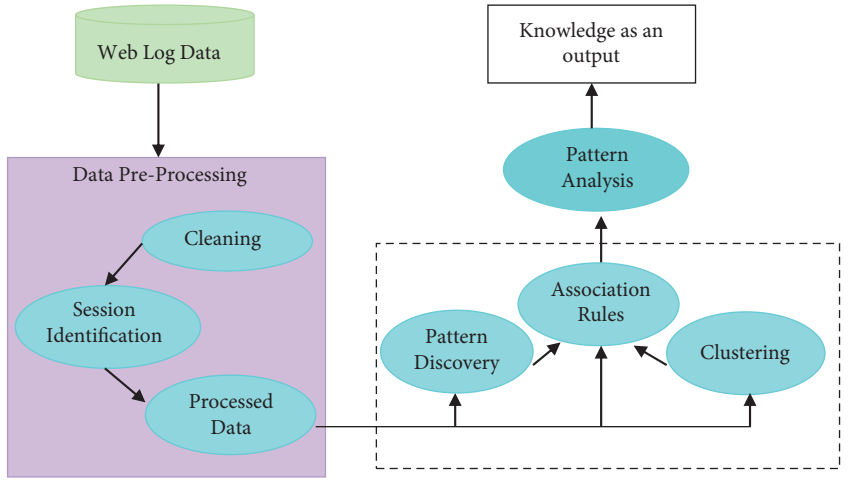

Figure 4: Architecture of web usages mining (WUM).

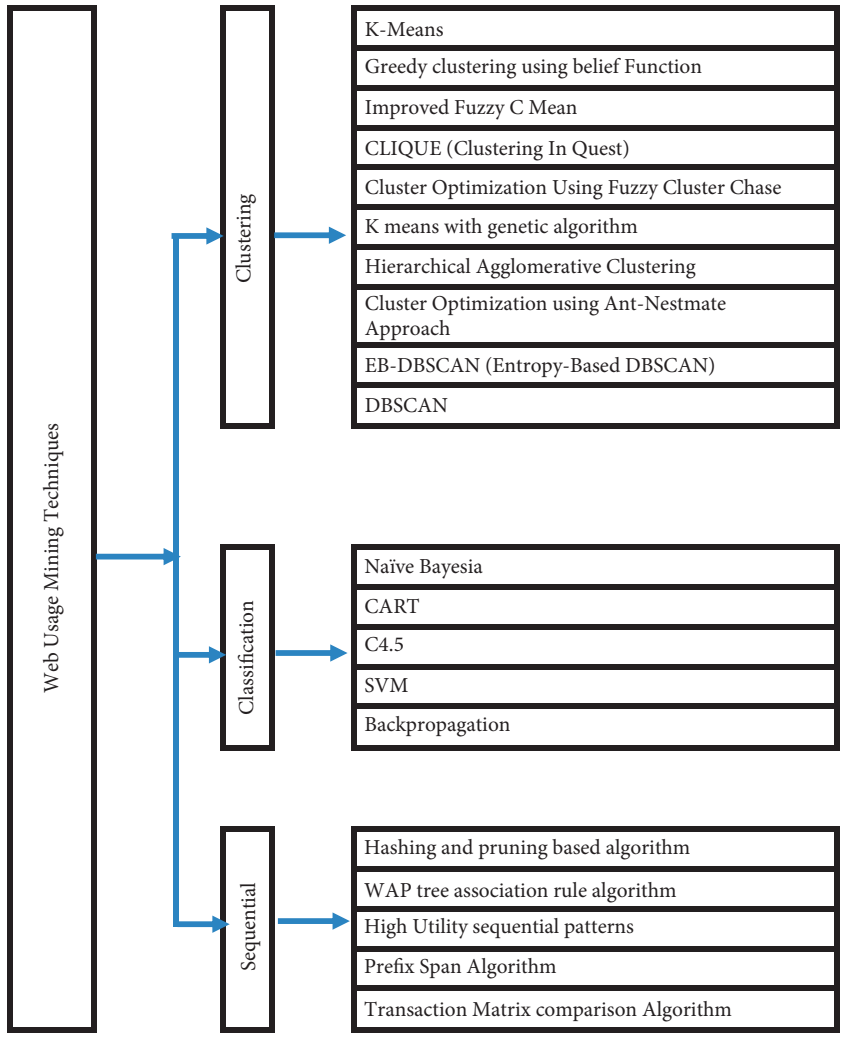

Figure 5: Algorithms used in web usage mining's techniques.

2.1.3. Clustering. Clustering is one of the most challenging unsupervised learning problems. Objects are sorted into groups of related members during the clustering process. As a result, a cluster is a group of things that are related to one another but not to the objects of other clusters. Clustering analysis is a method of grouping individuals or data objects (pages) with similar characteristics together. The formulation and execution of future marketing plans might be aided

by grouping user information or pages. The usage of user clustering will aid in the discovery of groups of users who have similar navigation patterns. Clustering techniques make sets of similar items from a large volume of data by using distance functions that compute the similarity ratio between items [30]. The contrast of the user/client and individual groups is an essential factor in such type of searching. There are two types of clustering available in this area:

(i) User clustering

(ii) Page clustering

User clustering is used to find those users who have the same browser patterns, and page clustering is used to find similar content's web pages.

2.1.4. Sequential Analysis. Sequential analysis is that which is found in those patterns in which one set or sets of pages are accessed one after another with a time sequence. For the prediction of future visitors, this application works by advertising on users group. Some techniques are utilized for sequential analysis [31], as shown in Figure 5. A detailed description of various algorithms of WUM Techniques is given in Table 2.

2.2. Web Content Mining (WCM). Web Content Mining (WCM) [13, 33, 34] as shown in Figure 6, is used to fetch relevant & Reliable information from web pages which may contain text documents, Hyperlinks, Structured data, audio, and Video. Nowadays, web pages are increasing exponentially over www.

Fetching relevant data related to user queries from an extensive collection of web pages is very difficult and time-consuming. Web content mining has the following approaches [33] to extract user relevant information from different types of data such as unstructured data, structure data, semistructured. There are various content mining algorithms [35] used by the above content mining techniques are shown in Table 3.

2.3. Web Structure Mining (WSM). Web Structure Mining detects the structural summary of a web page and its linked web pages as architecture shown in Figure 7. It finds out-link (forward/backwards) structure inside a web page by structure mining [33, 36]. It is used to classify and compare web documents and integrate the number of different web documents. Some of the popular Web structure mining algorithms are summarized in Table 4.

Web structure mining (WSM) as shown in Figure 7 follow the following steps:

(i) Apply link analysis on a web page repository to extract links (forward/backward) summary of web pages.

(ii) Apply a link mining techniques in the summary to find out the weight or quality of the web pages.

2.4. Challenges in Web Mining. Web mining is faced with some technical and nontechnical issues. Nontechnical issues occur due to management, fund, and resources (such as professional humans), Some technical issues are discussed below:

(i) *Inappropriate data:* Collected data should be reliable and in proper format to do successfully mining because many times data are incomplete and unavailable. It is very difficult to assure the accuracy of such a data.

(ii) *Complexity of web pages:* The structure of a web page is not predefined. It is stored in a digital library (order of data is not defined) in its original format. So, mining of data is very complex.

(iii) *Dynamic Web:* In dynamic web, data are frequently changed due to new updation. For example, sports data. Therefore, the complexity of mining is increased.

(iv) *Shortage of Mining Tools:* Need to develop a mining tool because a very smaller number of appropriate and complete mining tools is available.

2.5. Features of Web Page and Importance of These Features in a Ranking System. In this, we find out features of web pages and the importance of these features in the ranking system [29, 31], [30, 37–39] of the search engines (shown in Table 5). For each web page, there are fifteen features as given in the table, these features further divide into seven groups. All seven groups were finally categorized into three parts based on Web Mining types (WCM, WUM & WSM).

Page: It has two characteristics one of them is Page rank (PR) score and the second one is the age (AGE) of web pages in an index of search engine.

Links: It is associated with links/URLs (forward/Backward Links) on the web Page.

Query and Text Similarity: It indicates similarity ratio between query keywords and contents of a web page [40].

It has main three features:

(i) Frequency of query keywords inside title

(ii) Frequency of query keywords inside heading tags (H1, H2, . . ., H6) separately.

(iii) Frequency of query keywords inside paragraph.

Head Tag: Head tag contains two features: title and meta data. Both are used based on keywords inside title and meta description.

Body: it is associated with the density of keywords inside the body of a web page.

Content: associate with different features which are part of content analysis such as headings, links/URLs.

Session Specific: in this count total number of clicks, count unique clicks, and time duration for a session.

The above web parameters used in mining by Search Engine to find the quality and relevant web pages for

Table 2: Summary of algorithms of WUM techniques.

WUM techniques	Algorithms	Description
Clustering	K-means [8]	It is an unsupervised algorithm used for data mining and pattern recognition. The aim of K-Means is to minimize the cluster performance index
	Greedy clustering using belief function [10, 11]	It is used to modelling evidence from expert opinions or statistical information
	Improved fuzzy C mean [12]	It is a basic approach, used for image segmentation in which space divides into several clusters based on the pixel value of an image
	CLIQUE (clustering in quest) [13]	It is a subspace clustering algorithm that follows a bottom-up approach used to create static grids. This algorithm reduces the search space by using the apriori approach
	Cluster optimization using fuzzy cluster chase [14]	It is used to personalize web page clusters of end-users
	K means with genetic algorithm—minimizes objective function [15]	The GKA is the most preferable algorithm for clustering to other evolutionary algorithms
	Hierarchical agglomerative clustering [16]	It is a data exploratory analysis technique used in hierarchical clustering
	Cluster optimization using anti-estimate approach [17]	It is used to remove redundant data that may occur during clustering
	EB-DBSCAN (entropy-based DBSCAN) [18]	It is used to identify the high-density regions/areas
	DBSCAN [19]	It is used to make clusters of arbitrary shapes
Classification	Naïve Bayesia [20]	It is a work based on Bayes theorem to find a class with the highest probability from a predefined dataset by counting combination on values
	CART [21]	It is a classification technique used to construct decision trees for historical data
	C4.5 [22]	It is a quick classification & high precision algorithm. It is used frequently for classification
	SVM [32]	It is a classification algorithms that can be applied to linear and nonlinear datasets
	Backpropagation [23]	It is used as a gradient descent method to minimize error function in weight space
Sequential	Hashing and pruning based algorithm [24]	It is a famous association rule mining technique to increase the performance of traditional apriori algorithms
	WAP tree association rule algorithm [25]	WAP tree is a way to store the patterns in an effective manner by which these patterns are easily searchable
	High utility sequential patterns [26]	It is a data mining task that consists of a set of values having importance in a quantitative transaction database
	PrefixSpan algorithm [27]	It fetches sequential patterns using the pattern growth method. It works well for small datasets
	Transaction matrix comparison algorithm [28]	It uses a Boolean vector to discover frequent itemset. It required less memory because itemset stored in bits

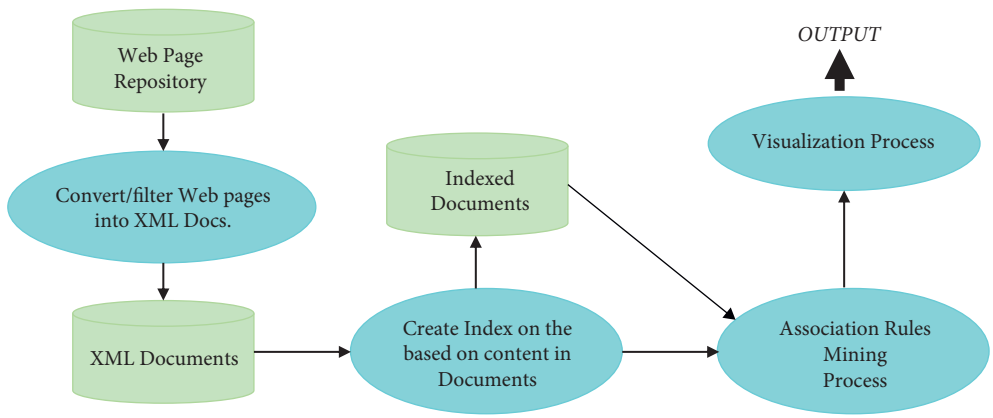

Figure 6: Architecture of web content mining (WCM).

Table 3: Summary of content mining algorithms.

Content mining algorithms	Description
Decision trees	It is a classification used by WCM and WUM. It is also a structured approach that contains root, branch, and leaf nodes. The root is split into subtrees/branches and the leaf contains a label of class
Naïve Bayes	It works based on Bayes theorem. To find a class with the highest probability from a predefined dataset by counting combinations of values. It is very powerful and easy to an understandable classifier
Support vector machine	It is a classification algorithm that can be applied to linear and nonlinear datasets. The separation of two classes (draw a decision boundary just as a line) depend on various classification features
Neural network	It works based on a backpropagation algorithm that contains an input layer, hidden layers, and an output layer. Each layer feeds to the next layer, and the number of hidden layers are arbitrary

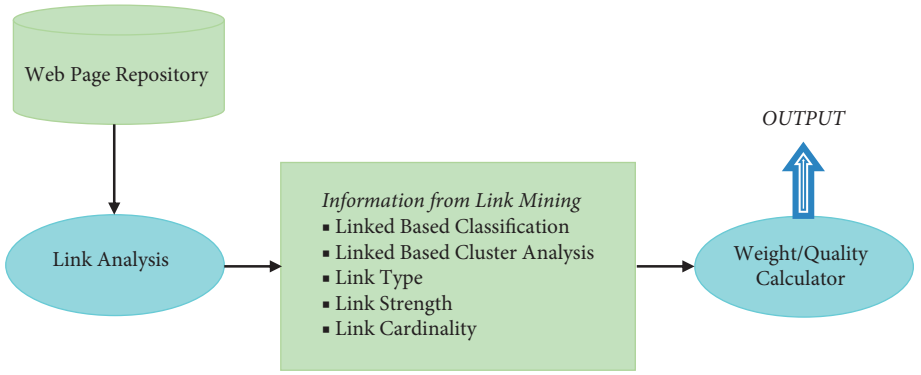

Figure 7: Architecture of web structure mining (WSM).

Table 4: Summary of structure mining algorithms.

Structure mining algorithm	Description
Page rank (PR)	Forward and backward links are used to compute the quality of a web page
Weighted page rank (WPR)	Compute the weight of the pages based on their structure (links) and this weight assigns to the page. Finally, generate rank based on weight
Eigen rumor (ER)	It is the modified version of WPR by applying some other parameters

Internet users for their queries. All the parameters are categorized according mining techniques. There are following web mining tools discussed in Table 6.

3. Web Page Ranking System

Every day, millions of people's access search engines to retrieve information according to their needs; hence, it becomes a common knowledge retrieval platform. The weight of the ranking in expert search for web documents is explained in [41]. The search engines have become the driver of Internet users that move them toward the highly ranked web by using various web mining techniques [42]. In order to maintain the ranking of web pages, the main objective of the website is to attract Internet users or clients so that they can maintain the ranking on renewed search engines. Reinforcement learning for Web Pages Ranking (WPR) algorithms is explained in [43]. There are several ways to improve the ranking of a web page on search engines, as SPAM farms are a very famous method to enhance a Website's ranking. During Rank calculation of web pages, cognitive spammer framework (CSF) deletes all spam web documents [44]. A framework Preference-based Universal Ranking Integration (PURI) [45] is designed by combining various ranking mechanisms. The Internet is an important source to access information from the web. At the same time, almost all web pages contain much noise such as advertisements, different types of banners, unreliable links that affect the performance of content and structure-based search engines, Question-Answering System, Web Summarization [13]. For instance, it fetches web documents based on the subject/title as given. To fetch huge web documents related to a specific domain is very easy and common. Therefore, to find reliable/matched web documents for user/client queries, search engines provide a ranking system. The g-index based expert-ranking system in which mainly Rep-FS, Exp-PC, and weighted Exp-PC techniques are used, explained in [46]. Ranking system utilize various web page ranking algorithms (as shown in Figure 8) like page rank [18, 47], weighted page rank [48], Eigenrumor [49], HITS [50], Weight Links Rank [21], distance ranking [51], tag rank [52], query dependent [17] to compute a rank of web page. It returns the order of web pages (order is done based on their rank).

Table 5: Summary of parameters used in mining by search engine.

Web mining techniques	Components of web page	Attributes	Description
*	Page	Rank	Ranking value of web page
		Age	Life of web page inside index of search engine
Web structure mining (WSM)	Links	Forward links	Number of links on that page point to other web pages
		Backward links	The number of web pages point to that page
Web content mining (WCM)	Query and text similarity	Freq_QK_Title	Number of query keywords in <title></title>
		Freq_QK_Heading	Number of query keywords in heading <h1>...<h6> tags
		Freq_QK_paragraph	Number of query keywords in paragraph tags
	Head Tag	Title	Keywords written inside <title></title>
		Metadata	Keywords in metadata key and description tag
	Body	Density	Keyword density
	Content	Heading	*
		Links	Heading keywords
			Images
			Paragraph
Web usage mining (WUM)	Session specific	Count clicks	Number of clicks during a session
		Count unique clicks	Number of unique click during a session
		Time duration for a session	Total time of a user session

Table 6: Summary of web mining tools.

Web mining tool	Mining used	Releasing year	Description
Scrapy	Web content mining (WCM)	26 June 2008	Scrapy is an open source and used to extract data from World wide web easily
Screen-scraper	Web content mining (WCM)	2001	It is used to find information from different webs for user queries. It used with proxy server mainly
Web content extractor	Web content mining (WCM)	January 5, 2012	It is used to fetch data from password protected web/hidden web
Mozenda	Web content mining (WCM)	—	It extract data from World Wide Web and send it to other different destinations
ORANGE	Web content mining (WCM)	2009	Used by ML (machine learning)
WEKA	Web usage mining (WUM)	1997	It is independent and open source
R	Web usage mining (WUM)	1997	Extract statistical data from graphical data
Website information filter system (SIFT)	Web usage mining (WUM)	—	It finds the patterns according user interest
Web utilization miner	Web usage mining (WUM)	—	It used to create a report of extracted data
Redwood	Web usage mining (WUM)	—	Web log is extracted by this mining tool

Page Rank is frequently used to calculate web page rank on the basis of in-link and out-link of the web page. The formula (shown in equation (2)) to calculate rank of a web page A

$$PageRank(A) = \sum_{B\epsilon Xa}^{\infty} \left(.\frac{\text{PageRank}(B)}{L(B)} \right). \quad (2)$$

Page rank of A is depend on the page rank value of each page B contained in the set of Xa (the set of all pages linking to page A), divided by number of links from page B.

L (B) -> Out-link from page B.

PageRank (B) -> Page rank of page B.

Weighted Page Rank is extended version of Page rank algorithm. It consider the popularity of web pages on the basis of link structure (in-links and out-links). WPR assign the different rank of the web page to its all the out-links.

Eigen Rumor is proposed to resolve the limitation of page rank and other web page ranking algorithm over blog i.e. it assign the rank value to each blog on the basis of weight of hub and authority of the blogger.

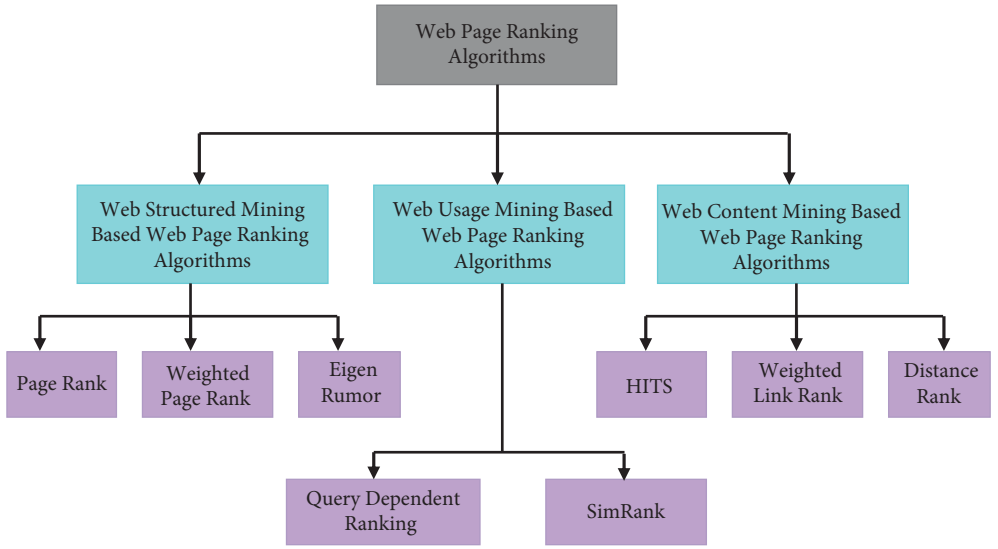

Figure 8: Web page ranking algorithms.

In query-dependent algorithm, use queries of the users to increase the performance of the page-ranking algorithm. A component was incorporated in the page-ranking algorithm which was dedicated to calculating the similarities between the user queries. The similarities between the user queries was analyzed by the algorithm and that information was used to decide the final results of return back to the user for a query.

A new approach (SimRank) using vector space model was proposed which uses the similarity from the vector space based model and finds the rank of the web page. The SimRank [17] algorithm assigns rank to the pages to be retrieved from the search engine in an effective way. Most of the traditional page rank algorithm uses the link structure of the web pages to find the page rank, and some of them are totally ignoring the content of the web pages. But SimRank algorithm also incorporates the content of the web pages to find the final rank score of a web page.

HITS algorithm computes rank of a web page by using popularity of web page. It also calculates the number of In-links and Out-links of a web page. The Hit based algorithm is basically computing the rank of a web page by calculating popularity of web page. The popularity is computes by determining input links (Authority) and output links (Hub) of a web page.

R. Baeza and E. Davis developed a Weighted Links Rank (WLR) algorithm with the help of standard PR algorithm. This algorithm generates weight of a link on the basis of three different arguments, that is, the anchors text length, tags, and relative position in the web page.

ZarehBidoki and Yazdani [14] proposed a reliable and intelligent web page ranking mechanism is called distance rank algorithm which is working on the basis of reinforcement learning algorithm. The distance between pages is calculated by using shortest logarithmic distance between 2-pages and assigns the rank accordingly to them. This algorithm returns very fast high-quality web pages by using distance based solution. For this algorithm, crawler takes more time to compute the distance vector for new web crawled web page. Table 7 shows the summary of web mining techniques and ranking algorithms for each mining technique.

4. State-of-the-Art Review

Due to increasing the information for humans on the WWW, the responsibility of the Internet also increased. It is straightforward for us to collect the information from www using search engines. Search engines return a large number of web pages as information for a user query. It is challenging for users to select reliable information among them. Therefore, in this section, we will discuss research papers in which the author tries to improve search engine techniques that support users to select reliable information.

In [54], authors give an approach to fetch experts' attributes by using text mining from the web, that is, it is a recommended model to return a precise record. This research has shown the effectiveness of the proposed approach in box-office revenue prediction. In [55], the author proposed a prediction for movie revenue based on YouTube trailer reviews. It is mainly utilized in business intelligence as well as in decision-making. In [56], the author developed a framework for Geographic Information Mining (GIM) framework. Microsoft discussion (MSD) forums used expert rank [57], a technique to find experts. This methodology used document-based relevance as well as authority. It does not consider MSD features (like rating by the user, which is a more reliable feature used to mine expert users). In [58], author identified user activities in the SO-forum and compared them with their GitHub repositories and feasible features of the user (active in both platforms). In [59], author proposed user activity models for stack overflow, Wenwo Forums & SinaWeibo to classify real experts. In [60], the model uses some basic features to compute the user weight. In this model, the question-answer ratio is used to generate user weight; still, it ignores the consistency of the user.

Table 7: Summary of web mining techniques based on various parameters.

Web mining techniques	Data	Algorithms	Methodology	List of input parameters	Complexity	Relevancy	Shortcomings
Web structured mining (WSM) [53]	Hyperlinks, structure of documents	Page rank (PR)	Forward and backward links are used to compute the quality of the web pages	Forward and backward links	PR take O (log Time to compute the rank	Return more relevant web pages	It is not considered the content of the web page
		Weighted page rank (WPR)	Compute the weight of the pages based on their structure (links) and assign to the page. Finally, generate rank based on weight	Forward and backward links	The time complexity of WPR is O (log N)	It returns more relevant web pages as compared to the page rank algorithm	It is not considered the content of the web page
		Eigen rumor (ER)	It is the modified version of WPR by applying some other parameters	Blogs, forward and backward links	The time complexity of ER is log N	It provides more relevant as compare to PR & WPR	Rely on web structure to compute page rank
Web content mining (WCM) [53]	Text, image, audio, video, structure record	HITS	Monitor and consider those web pages which are visited by Internet users regularly	History of users' log files	Time complexity is O (log N)	The relevancy of this approach is moderate	Due to only rely on it is not much efficient
		Weight links rank (WL rank)	It use the position of forward/ backward links to compute the rank of web page	Link structure and content	—	—	It is not reliable because links are not placed at the proper location/position
		Distance ranking (DR)	The reinforcement learning algorithm is used by DR to compute the rank of the web page	After crawling, it computes the distance between 2-web pages	Time taken by DR is O (log N)	—	The logarithmic distance between 2-web pages is not reliable
		Tag rank (TR)	Hub & authority concept are used to generate the rank	Forward, backward and tags on web pages	TR takes O (log N) time	Relevancy is moderate	It is less efficient
		Dirichlet rank	It uses attributes of link like position, anchor text, tag where links exist, to compute the rank	Position of link, name of tags in which link exist, anchor text	—	—	The relevancy is not affected by the position of the link
Web usage mining (WUM) [53]	Web server logs, app server logs, app level logs	Query dependent ranking	It uses similar user queries to compute the rank	Training queries	—	Relevancy is high. It returns all the same web pages to a same user query	There are limited numbers of characteristics to find the similarity

Table 8: Summary of previous research on basis of various parameters.

S. No.	Year	Title	Journal	Author	Methodology/approach	Advantages	Limitations
1	2021	Using machine learning for web page classification in search Engine optimization [92]	Future internet 2021, 13, 9	Goran Matosevi, Jasminka Dobsa, Dunja Mladeni	Authors used machine learning to classify web pages in SEO	Methods used in this research can help in building automated or semiautomated software for supporting SEO work	It is language-specific
2	2021	Learning to rank for educational search engines [29]	IEEE transactions	Arif Usta, Ismail Sengor Altingovde, Rifat Ozcan, and Ozgur Ulusoy	Machine learning (ML) techniques. It is also called learning to Rank (LTR)	It exposes general and user query-dependent ranking models. It used LTR to trained, to increase high reliability in educational search. It provides better learning practice	It has domain-specific features, and increases the perceived latency due to query dependence
3	2020	Big data analytics for search engine optimization [93]	MDPI, big data and cognitive computing	Ioannis C. Drivas, Damianos P. Sakas, Georgios A. Giannakopoulos, Daphne Kyriaki-Manessi	Author used agent based model, fuzzy cognitive mapping and big data analytics	It increases the organic search engine visits by using multiple SEO factors	This process of search engine optimization (SEO) could be a cost-effective
4	2020	Incremental refinement of page ranking of web pages [94]	*International Journal of Information Retrieval Research*, vol. 11, no. 2	P. S. Sharma, Divakar Yadav	Author used frequency of query keyword, hyperlink on query keywords and proxy server approach	It improves web page ranking and reduces perceived latency	Authors applied this approach only on web data
5	2019	Natural-language-based intelligent retrieval engine for BIM object database [95]	Computers in industry (Elsevier)	Songfei Wua, Qiyu Shena, Yichuan Denga, Jack Chengc	NLP (natural language processing)	It reduce threshold by using BIM object database	It is domain specific. Used ontology to understand semantic
6	2019	Jail-phish: An improved search engine-based phishing detection system [96]	Computers in industry (Elsevier)	Routhu Srinivasa Rao, Alwyn Roshan Pais	Author used heuristic technique to fetch similarity-based features to identify the phishing websites	It is used to identify the malicious users and also to detect phishing sites. It also works on free hosted websites	It is used as third-party based features to identify phishing sites on free hosted webs
7		Forecasting tourist arrivals with machine learning and Internet search index [97]	Tourism management (Elsevier)	Shaolong Suna, Yunjie Weia, Kwok-Leung Tsuic, Shouyang Wanga	Author used Machine learning (ML) and indexes of search engines	It increase forecasting accuracy and robustness	It is tested on 1- test case. It works on keyword selection
8	2019	An investigation of biases in web search engine query suggestions [98]	Online information review, Vol. 44 no. 2, 2020, pp. 365–381 © Emerald Publishing Limited	Malte Bonart, Anastasiia Samokhina, Gernot Heisenberg and Philipp Schaer	Author designed a framework that automatically analyzes query suggestions for the web user	It is capable of Automatically collecting and analyzing query suggestions for a large repository of search keywords	It is topics derived. It is for the politician domain only
9	2018	Improving search engine optimization (SEO) by using hybrid Modified MCDMmodels [99]	Artificial intelligence review https://doi.org/10.1007/s10462-018-9644-0 (Springer)	Hung-Jia Tsuei, Wei-Ho Tsai, Fu-Te Pan, Gwo-Hshiung Tzeng	Multi-criteria decision-making (MCDM, also known as multicriteria decision analysis, MCDA)	Improving and evaluating search engine ranking	Need to improve on low-value websites

Table 8: Continued.

S. No.	Year	Title	Journal	Author	Methodology/approach	Advantages	Limitations
10	2017	IBRI-CASONTO: Ontology-based semantic search engine [100]	Egyptian informatics journal (Elsevier)	Awny Sayed, Amal Al Muqrishi	Resource description framework (RDF) data & ontological graph	It supports Arabic and English language. It use keyword-based search and a semantics-based search	It is domain specific. Indexing mechanism is not working for large data sets
11	2017	Death prediction and analysis using web mining techniques [101]	ICACCS - 2015, Coimbatore, India-978-1-5090-4559–4/17/$31-EEE	Hesham Abdo, Ahmed Aqlan, et al.	Author used regression and neural network-based methodology to compute the rank of the webpages	It used the latest AI methodologies to improve page rank algorithms	Need to find suitable AI techniques to improve the prediction approach
12	2017	Design of a framework for knowledge based web page ranking [102]	International Journal of Engineering and Technology	P. Sharma, S. A.K, and P. Garg	Author used frequency of query keywords and proxy server approach	It reduce perceived latency	Synchronization problem between proxy and search engines
13	2016	New query suggestion framework and algorithms: A case study for an educational search engine [103]	Information processing and management (elsevier)	Bahattin Vidinli, Rifat Ozcan	Authors designed a framework query suggestion	It can be reduced (simplified) to a problem of query compression	Need to be including spell checker to increase suggestion accuracy and needs to integrate content also
14	2015	Search-based QoS ranking prediction for web services in cloud environments [104]	Future generation computer systems (Elsevier)	Mao, Chengying Chen, Jifu Towey, Dave Chen, Jinfu Xie, Xiaoyuan	Author explore similarity measurement method for two ranked sequences. It predicts QoS ranking	The QoS information was used for ranking prediction	It does not use social relationships in the cloud platform which is important
15	2014	Effective ranking and search techniques for web resources considering semantic relationships [105]	Information processing and management (elsevier)	Lee, Jihyun Min, Jun Ki Oh, Alice Chung, Chin Wan	Use ontology to compute weight for the semantic relationship	It increased the power of the query keyword for semantic relationship. It reduced the search space and increased the accuracy	It used ontology
16	2013	A hybrid approach for extracting informative content from web pages [33]	Information processing and management 49 (2013) 928–944 contents (Elsevier)	Erdinç Uzun, Hayri Volkan Agun, Tarık Yerlikaya	It uses decision tree learning to fetch informative contents and make rules. Extract reliable information by using these rules	It is very faster after making rules and provides high accuracy in results	It takes more time for rules creation the first time
17	2013	Topic-Driven SocialRank: Personalized search result ranking by identifying similar, credible users in a social network [35]	Knowledge-based systems 54 (2013) 230–242 contents (elsevier)	Young An Kim, Gun Woo Park	Focus on identifying similar users who have high credibility and sharing their search experiences	It is very useful to find more relevant search results by implicit help of familiar, credible users	It is tested on a small dataset
18	2012	Mining the real-time web: A novel approach to product recommendation [34]	Knowledge-based systems 29 (2012) 3–11 (Elsevier)	Garcia Esparza, Sandra O'Mahony, Michael P. Smyth, Barry	It uses a collaborative-filtering based approach	It is used for micro-blogging messages	It is not used by other domains like Twitter
19	2011	A music information system automatically generated via web content mining techniques Markus [36]	Information processing and management 47 (2011) 426–439 (Elsevier)	Markus Schedl, Gerhard Widmer, Peter knees, Tim Pohle	It uses web content mining techniques	It provides web-based access to a large collection of music artists. It is automated music information system.	It is domain specific (for music only)
20	2011	Snoogle: A search engine for pervasive environments [106]	IEEE transactions on parallel and distributed systems	Haodong Wang, Chiu C. Tan, and Qun Li Abstract—Embedding	It uses sensor networks, And communication overhead reduced by Bloom filters	In this, user can search a mobile object (s) that fit in detail	This system is not able to find a moving object in real-time

Besides this, the quality of the tag was not considered. Although, it may lead to more reliable and accurate recommendation systems. The link-based expert finding techniques mainly used the structure of links instead of their contents. Link analysis used question-answer relationship [61], to find experts, citation networks [62] and e-mail communications [63]. For online users, in [64], the author presented an automatic expert-finding model. In this model, the profile of user expertise was evaluated based on social network score and postconditions. The Z-Score, PageRank, In-degree, and HITS, etc., algorithms were used to compute social network authority scores. A search engine to fetch biomedical information [65] return all the documents corresponding user query from MEDLINE based on word/concept indexes. Several researchers have investigated various ranking approaches by using different methodologies that increase the efficiency of search engines to provide highly relevant web pages for a particular user query.

In [66], recent research in CARSs is mainly directed by developing novel techniques or adapting and combining existing ones that can efficiently deal with the growing complexity and dynamicity of social networks.

The main consequences of the [67] are (1) ontologies of a corpus can be organized effectively, (2) no effort and time are required to select an appropriate ontology, (3) computational complexity is only limited to the use of unsupervised learning techniques, and (4) due to no requirement of context awareness, the proposed framework can be effective for any.

In [68] author explores the various explanation techniques to identify the local contribution of ranking indicators based on the position of an instance in the ranking as well as the size of the neighborhood around the instance of interest. We evaluate the generated explanations for the Times Higher Education University ranking dataset as a benchmark of competitive ranking.

Table 8 summarizes various research papers [69–85] based on different attributes such as methodologies, approaches, pros and cons, etc. Additionally, futuristic research directions in similar areas are presented in [86–91].

4.1. Observations from the State-of-the-Art- Reviews. Following observation are made after the critical review of the state-of-the-art review:

Observation 1: Mostly search engines return relevant web pages to users for their queries. Relevancy of web page depends upon in-link/ out–link (i.e., web structure mining) and popularity of web page.

Many times, the most relevant web pages may be less important for user queries. Important web pages, according to user queries may be missing out from the result. So new techniques are required to develop that may consider user queries as an additional parameter to find the relevant web pages for those queries.

Observation 2: Due to increasing the size of the web, search engines delay returning a list of web pages as output to users. The delay between user query submissions and to get output is called perceived latency. Therefore, a pre-fetching mechanism needs to be developed to reduce the response time.

Observation 3: Even with the introduction of a prefetching mechanism that aims to reduce the user perceived latency, unsuccessful predictions made to prefetch the pages may result in information overkill. Thus, a mechanism is required that could actually make credible predictions for only those pages that are more relevant, that is, make correct predictions to minimize the problem of information overkill.

Observation 4: Due to increasing WWW and Internet users, it is very difficult to fetch the information, which is looked at, by a specific group of users. For example, in an organization all employees may request the same type of information. Therefore, it requires approaches that personalize the content of web pages with respect to the user's group.

A critical look at the available state-of-the-art reviews reveals that the following major gaps are identified:

(1) Possibility of existence important page but less popular, which may not be linked

(2) Delay in response as perceived by user

(3) Need to search information in a similar interest group in an organization

5. Conclusion and Future Scope

Three categories of ranking algorithms are mainly discussed. The first category of algorithm based on the content of web pages is known as content-based page ranking. The second category of the algorithm, which uses the link structure of the worldwide web, is known as web structure-based page ranking algorithms, and the third category used a hybrid of the first and second categories. Ranking systems highly rely upon web mining techniques, but some issues need to be addressed in web mining due to improper data, shortage of mining tools, and other challenges in classification and clustering techniques.

The existing ranking systems have several limitations, which define the challenge and new research paths for researchers. The observations about existing research work will help the researcher select the specific area where further research may be initiated.

There are some challenges related to web page ranking, such as the following:

(i) Web structure-based page ranking algorithms may ignore web pages with less page ranking score but good content for a user query. Content-based page ranking algorithms take more time to find page rank because of content mining at query time.

(ii) The size of WWW is huge, so content mining is a very time-consuming process to check the quality of web pages. There is a need to reduce the time taken by search engines to return the results.

(iii) To improve the search results for user queries, it is needed to search for information in a similar interest group in an organization [107–113].

Conflicts of Interest

There are no conflicts of interest.

Acknowledgments

This research was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R195), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important factors for improving Google search rank," *Future Internet*, vol. 11, no. 2, p. 32, 2019.

[2] C. Fu, C. Peng, X.-Y. Liu, L. T. Yang, J. Yang, and L. Han, "Search engine: the social relationship driving power of Internet of Things," *Future Generation Computer Systems*, vol. 92, pp. 972–986, 2019.

[3] D. Qiao, J. Zhang, Q. Wei, and G. Chen, "Finding competitive keywords from query logs to enhance search engine advertising," *Information & Management*, vol. 54, no. 4, pp. 531–543, 2017.

[4] M. Maheshwari and R. Ali, "Evolution of search engine optimization and investigating the effect of panda update into it," *International Journal of Scientific & Engineering Research*, vol. 4, no. 12, pp. 2045–2053, 2013.

[5] J. V. Glowniak, ""Information retrieval on the internet and medical resources on the internet," *Annals of Internal Medicine*, vol. 123, no. 2, pp. 127–131, 1995.

[6] Z. Hadjilambrou, M. Kleanthous, G. Antoniou, A. Portero, and Y. Sazeides, "Comprehensive characterization of an open source document search engine," *ACM Transactions on Architecture and Code Optimization*, vol. 16, no. 2, pp. 1–21, 2019.

[7] K. Mohan, "A survey on web structure mining," *International Journal of Advanced Computer Research*, vol. 1, no. 1, pp. 715–720, 2017.

[8] Cisco, "CISCO: global - 2021 forecast highlights," https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf.

[9] M. Arora, U. Kanjilal, and D. Varshney, "Evaluation of information retrieval: precision and recall," *International Journal of Indian Culture and Business Management*, vol. 12, no. 2, p. 224, 2016.

[10] N. Höchstötter and D. Lewandowski, "What users see - structures in search engine results pages," *Information Sciences*, vol. 179, no. 12, pp. 1796–1812, 2009.

[11] H.-J. Tsuei, W.-H. Tsai, F.-T. Pan, and G.-H. Tzeng, "Improving search engine optimization (SEO) by using hybrid modified MCDM models," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 1–16, 2020.

[12] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing & Management*, vol. 42, no. 1, pp. 248–263, 2006.

[13] E. Uzun, H. V. Agun, and T. Yerlikaya, "A hybrid approach for extracting informative content from web pages," *Information Processing & Management*, vol. 49, no. 4, pp. 928–944, 2013.

[14] A. Usta, I. S. Altingovde, R. Ozcan, and O. Ulusoy, "Learning to rank for educational search engines," *IEEE Transactions on Learning Technologies*, vol. 14, no. 2, pp. 211–225, 2021.

[15] O. Dan and B. D. Davison, "Measuring and predicting search engine users' satisfaction," *ACM Computing Surveys*, vol. 49, no. 1, pp. 1–35, 2016.

[16] C.-J. Luh, S.-A. Yang, and T.-L. D. Huang, "Estimating Google's search engine ranking function from a search engine optimization perspective," *Online Information Review*, vol. 40, no. 2, pp. 239–255, 2016.

[17] L.-W. Lee, J.-Y. Jiang, C. Wu, and S.-J. Lee, "A query-dependent ranking approach for search engines," in *Proceedings of the 2009 Second International Workshop on Computer Science and Engineering*, vol. 1, pp. 259–263, Qingdao, China, October 2009.

[18] R. Baeza-Yates and E. Davis, "Web page ranking using link attributes categories and subject descriptors," in *Proc. 13th Int. World Wide Web Conf. Altern. track Pap. posters*, pp. 328-329, New York NY USA, May 2004.

[19] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G. V. Uma, and K. Sarukesi, "Relevance ranking and evaluation of search results through web content mining," *Lecture Notes in Engineering and Computer Science*, vol. 2195, pp. 456–460, 2012.

[20] R. K. Roul, S. R. Asthana, and G. Kumar, "Spam web page detection using combined content and link features," *International Journal of Data Mining, Modelling and Management*, vol. 8, no. 3, pp. 209–222, 2016.

[21] P. Sharma, D. Tyagi, and P. Bhadana, "Weighted page content rank for ordering web search result," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7301–7310, 2010.

[22] P. S. Sharma and D. Yadav, "Incremental refinement of page ranking of web pages," *International Journal of Information Retrieval Research*, vol. 10, no. 3, pp. 57–73, 2020.

[23] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 977–990, 2011.

[24] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker, "Searching and browsing linked data with SWSE: the semantic web search engine," *Journal of Web Semantics*, vol. 9, no. 4, pp. 365–401, 2011.

[25] M. Jawad and H. Mughal, "Data mining : web data mining techniques, tools and algorithms : an overview," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, pp. 208–215, 2018.

[26] A. B. Can and N. Baykal, ""MedicoPort: a medical search engine for all," Comput," *Methods Programs Biomed*, vol. 86, no. 1, pp. 73–86, 2007.

[27] A. Anagnostopoulos, A. Z. Broder, and D. Carmel, "Sampling search-engine results," *World Wide Web*, vol. 9, no. 4, pp. 397–429, 2006.

[28] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Computer Networks*, vol. 50, no. 10, pp. 1448–1463, 2006.

[29] M. Zorrilla and D. García-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners," *Decision Support Systems*, vol. 55, no. 1, pp. 399–411, 2013.

[30] M. Kumari and S. Soni, "A review of classification in web usage mining using K- nearest neighbour," *Advances in Computational Sciences and Technology*, vol. 10, no. 5, pp. 1405–1416, 2017.

[31] P. Suthar and P. B. Oza, "A survey on web usage mining techniques," *International Journal of Computer Science and Information Technology*, vol. 2, no. 10, pp. 3824–3829, 2015.

[32] T. Seymour, D. Frantsvog, and S. Kumar, "History of search engines," *International Journal of Management & Information Systems*, vol. 15, no. 4, p. 47, 2011.

[33] F. Johnson and S. Kumar Gupta, "Web content mining techniques: a survey," *International Journal of Computer Application*, vol. 47, no. 11, pp. 44–50, 2012.

[34] A. kumar, "A study on web content mining," *International Journal Of Engineering And Computer Science*, vol. 6, no. 1, pp. 2015–2018, 2017.

[35] R. Malarvizhi and K. Saraswathi, "Web content mining techniques tools & algorithms – a comprehensive study," *Web Content Min. Tech. Tools Algorithms – A Compr. Study*, vol. 4, no. 8, pp. 2940–2945, 2013.

[36] C. Hahnel, F. Goldhammer, U. Kröhne, and J. Naumann, "The role of reading skills in the evaluation of online information gathered from search engine environments," *Computers in Human Behavior*, vol. 78, pp. 223–234, 2018.

[37] C. Vyas, "Evaluating state tourism websites using Search Engine Optimization tools," *Tourism Management*, vol. 73, no. January, pp. 64–70, 2019.

[38] Y. Li and H. Wu, "A clustering method based on K-means algorithm," *Physics Procedia*, vol. 25, pp. 1104–1109, 2012.

[39] T. Denoeux, *Methods for Building Belief Functions*, pp. 1–67, University of Technology of Compiégne, Compiégne, France, October 2019.

[40] A. Bakhthemmat and M. Izadi, "Communities detection for advertising by futuristic greedy method with clustering approach," *Big Data*, vol. 9, no. 1, pp. 22–40, 2021.

[41] C. MacDonald and I. Ounis, "The influence of the document ranking in expert search," *Information Processing & Management*, vol. 47, no. 3, pp. 376–390, 2011.

[42] K. C. Lee and S. Lee, "Interpreting the web-mining results by cognitive map and association rule approach," *Information Processing & Management*, vol. 47, no. 4, pp. 482–490, 2011.

[43] V. Derhami, E. Khodadadian, M. Ghasemzadeh, and A. M. Zareh Bidoki, "Applying reinforcement learning for web pages ranking algorithms," *Applied Soft Computing*, vol. 13, no. 4, pp. 1686–1692, 2013.

[44] A. Makkar and N. Kumar, "An efficient deep learning-based scheme for web spam detection in IoT environment," *Future Generation Computer Systems*, vol. 108, pp. 467–487, 2020.

[45] J. M. García, M. Junghans, D. Ruiz, S. Agarwal, and A. Ruiz-Cortés, "Integrating semantic Web services ranking mechanisms using a common preference model," *Knowledge-Based Systems*, vol. 49, pp. 22–36, 2013.

[46] M. S. Faisal, A. Daud, A. U. Akram, R. A. Abbasi, N. R. Aljohani, and I. Mehmood, "Expert ranking techniques for online rated forums," *Computers in Human Behavior*, vol. 100, no. December, pp. 168–176, 2019.

[47] T. W. L. Page, S. Brin, and R. Motavni, "The PageRank citation ranking: bringing order to the web," Tech. Report, Stanford Digit. Libr. Technol, Stanford InfoLab, Stanford, 1998.

[48] W. Xing and A. Ghorbani, "Weighted PageRank algorithm," in *Proceedings of the - Second Annu. Conf. Commun. Networks Serv. Res.*, pp. 305–314, Fredericton, NB, Canada, May 2004.

[49] K. Fujimura, N. Tanimoto, and M. Sugisaki, "The eigenrumor algorithm for ranking blogs," in *Proceedings of the Second Annu. Work. Weblogging Ecosyst. Aggregation*, pp. 59–74, Chiba, Japan, July 2005.

[50] J. K. S. Chakrabarti, B. E. Dom, S. R. Kumar et al., "Mining the link structure of the World Wide web," *Computer*, vol. 12, no. 23, pp. 60–67, 1999.

[51] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: an intelligent ranking algorithm for web pages," *Information Processing & Management*, vol. 44, no. 2, pp. 877–892, 2008.

[52] S. Jie, C. Chen, Z. Y. Rong-Shuang et al., "TagRank: a new rank algorithm for webpage based on social web," in *Proceedings of the 2008 International Conference on Computer Science and Information Technology*, pp. 254–258, Singapore, September 2008.

[53] H. Rustum, H. Hadi, and A. AbdulZahraa, "Improved fuzzy C-mean algorithm for image segmentation," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 6, 2016.

[54] X. Xie, Y. Fu, H. Jin, Y. Zhao, and W. Cao, "A novel text mining approach for scholar information extraction from web content in Chinese," *Future Generation Computer Systems*, vol. 111, pp. 859–872, 2020.

[55] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie revenue prediction based on purchase intention mining using YouTube trailer reviews," *Information Processing & Management*, vol. 57, no. 5, Article ID 102278, 2020.

[56] Y. Zhang, Q. Ma, Y.-Y. Chiang et al., "Extracting geographic features from the Internet: a geographic information mining framework," *Knowledge-Based Systems*, vol. 174, pp. 57–72, 2019.

[57] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "ExpertRank: a topic-aware expert finding algorithm for online knowledge communities," *Decision Support Systems*, vol. 54, no. 3, pp. 1442–1451, 2013.

[58] B. Vasilescu, V. Filkov, A. Serebrenik, and A. Serebrenik, "StackOverflow and GitHub: associations between software development and crowdsourced knowledge," in *Proceedings of the 2013 International Conference on Social Computing*, Alexandria, VA, USA, September 2013.

[59] Z. Liu and B. J. Jansen, "Questioner or question: p," *Information Processing & Management*, vol. 54, no. 2, pp. 159–174, 2018.

[60] I. Cantador, A. Bellogín, M. E. Cortés-Cediel, O. Gil, and O. Gil, "Personalized recommendations in e-participation : offline experiments for the 'decide madrid' platform," in *Proceedings of the International Workshop on Recommender Systems for Citizens*, Como Italy, August 2017.

[61] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert finding for question answering via graph regularized matrix completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 993–1004, 2015.

[62] S. Divya and R. . Mall, "Structural analysis of L-bracket using ansys," *i-manager's Journal on Mechanical Engineering*, vol. 7, no. 2, p. 17, 2017.

[63] H. Kareem and L. Asker, "Detecting hierarchical ties using link-analysis ranking at different levels of time granularity," pp. 1–4, https://arxiv.org/abs/1701.06861.

[64] J. Zhang, M. S. Ackerman, and K. K. Nam, "QuME : a mechanism to support expertise finding in online help-seeking communities," in *Proceedings of the 20th annual ACM symposium on User interface software and tech- nology*, pp. 111–114, Newport Rhode Island USA, October 2007.

[65] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "FACTA: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559-2560, 2008.

[66] A. B. Suhaim and J. Berri, "Context-Aware recommender systems for social networks: review, challenges and opportunities," *IEEE Access*, vol. 9, pp. 57440–57463, 2021.

[67] M. A. Sarwar, M. Ahmed, A. Habib et al., "Exploiting ontology recommendation using text categorization approach," *IEEE Access*, vol. 9, pp. 27304–27322, 2021.

[68] H. Anahideh and N. Mohabbati-Kalejahi, "Local explanations of global rankings: insights for competitive rankings," *IEEE Access*, vol. 10, pp. 30676–30693, 2022.

[69] G. Matošević, J. Dobša, and D. Mladenić, "Using machine learning for web page classification in search engine optimization," *Future Internet*, vol. 13, no. 1, pp. 9–20, 2021.

[70] I. C. Drivas, D. P. Sakas, G. A. Giannakopoulos, and D. Kyriaki-Manessi, "Big data analytics for search engine optimization," *Big Data and Cognitive Computing*, vol. 4, no. 2, pp. 5–22, 2020.

[71] S. Wu, Q. Shen, Y. Deng, and J. Cheng, "Natural-language-based intelligent retrieval engine for BIM object database," *Computers in Industry*, vol. 108, pp. 73–88, 2019.

[72] R. S. Rao and A. R. Pais, "Jail-Phish: an improved search engine based phishing detection system," *Computers & Security*, vol. 83, pp. 246–267, 2019.

[73] S. Sun, Y. Wei, K.-L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," *Tourism Management*, vol. 70, pp. 1–10, July 2019.

[74] M. Bonart, A. Samokhina, G. Heisenberg, and P. Schaer, "An investigation of biases in web search engine query suggestions," *Online Information Review*, vol. 44, no. 2, pp. 365–381, 2019.

[75] A. Sayed and A. Al Muqrishi, "IBRI-CASONTO: ontology-based semantic search engine," *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 181–192, 2017.

[76] A. D. Hesham Abdo Ahmed Aqlan and S. Ahmed, "Death prediction and analysis using web mining techniques," in *Proceedings of the 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, January 2017.

[77] P. Sharma, A. K. Sharma, and P. Garg, "Design of a framework for knowledge based web page ranking," *International Journal of Engineering & Technology*, vol. 9, no. 3, pp. 2236–2244, 2017.

[78] I. B. Vidinli and R. Ozcan, "New query suggestion framework and algorithms: a case study for an educational search engine," *Information Processing & Management*, vol. 52, no. 5, pp. 733–752, 2016.

[79] C. Mao, J. Chen, D. Towey, J. Chen, and X. Xie, "Search-based QoS ranking prediction for web services in cloud environments," *Future Generation Computer Systems*, vol. 50, pp. 111–126, 2015.

[80] J. Lee, J.-K. Min, A. Oh, and C.-W. Chung, "Effective ranking and search techniques for Web resources considering semantic relationships," *Information Processing & Management*, vol. 50, no. 1, pp. 132–155, 2014.

[81] Y. A. Kim and G. W. Park, "Topic-Driven SocialRank: personalized search result ranking by identifying similar, credible users in a social network," *Knowledge-Based Systems*, vol. 54, pp. 230–242, 2013.

[82] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth, "Mining the real-time web: a novel approach to product recommendation," *Knowledge-Based Systems*, vol. 29, pp. 3–11, 2012.

[83] M. Schedl, G. Widmer, P. Knees, and T. Pohle, "A music information system automatically generated via Web content mining techniques," *Information Processing & Management*, vol. 47, no. 3, pp. 426–439, 2011.

[84] H. Haodong Wang, C. C. Tan, and Q. Qun Li, "Snoogle: a search engine for pervasive environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 8, pp. 1188–1202, 2010.

[85] S. Bhatia and A. Tyagi, "Twitter Trends reveals: focus of interest in the sleep trend analytics on response to COVID-19 outbreak," *Current Psychiatry Research and Reviews*, vol. 16, p. 1, 2020.

[86] A. Ahmed, A. R. Javed, Z. Jalil, G. Srivastava, and T. R. Gadekallu, "October. Privacy of web browsers: a challenge in digital forensics," in *International Conference on Genetic and Evolutionary Computing*, pp. 493–504, Springer, Singapore, 2021.

[87] M. Bhatia, S. Sharma, S. Bhatia, and M. Alojail, "Fog computing mitigate limitations of cloud computing," *International Journal of Recent Technology and Engineering (IJRTE) ISSN*, vol. 8, no. 5, pp. 2277–3878, 2019.

[88] S. Bhatia, M. Sharma, K. K. Bhatia, and P. Das, "Opinion target extraction with sentiment analysis," *International Journal of Computing*, vol. 17, no. 3, pp. 136–142, 2018, Q3.

[89] S. Bhatia, M. Sharma, M. Sharma, and K. K. Bhatia, "Opinion score mining: an algorithmic approach," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 11, pp. 34–41, 2017.

[90] S. Bhatia, M. Sharma, and K. K. Bhatia, "A novel approach for crawling the opinions from World Wide Web," *International Journal of Information Retrieval Research*, vol. 6, no. 2, pp. 1–23, 2016.

[91] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[92] P. Contreras and F. Murtagh, "Hierarchical clustering," *Handb. Clust. Anal., no. February*, pp. 103–124, 2015.

[93] A. Alphy and S. Prabakaran, "Cluster optimization for improved web usage mining using ant nestmate approach," in *Proceedings of the 2011 International Conference on Recent Trends in Information Technology (ICRTIT) 2011*, pp. 1271–1276, Chennai, India, June 2011.

[94] T. Wang, C. Ren, Y. Luo, and J. Tian, "Ns-Dbscan: NS-DBSCAN: a density-based clustering algorithm in network space," *ISPRS International Journal of Geo-Information*, vol. 8, no. 5, pp. 218–5, 2019.

[95] M. Daszykowski and B. Walczak, "Density-based clustering methods," *Comprehensive Chemometrics*, vol. 2, pp. 635–654, 2009.

[96] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1–3, pp. 403–412, January, 2018.

[97] D. Steinberg, *CART: Classification and Regression Trees*, Taylor & Francis, no. January 2009, Oxfordshire, 2015.

[98] S. Idriss and A. Lawan, "An improved C4.5 model classification algorithm based on taylor's series," *Jordanian Journal of Computers and Information Technology*, vol. 5, no. 1, pp. 1–42, 2019.

[99] M. Awad and R. Khanna, "Machine learning and knowledge discovery," in *Efficient Learning Machines*, Apress, Berkeley, CA, USA, 2015.

[100] I. Jung and G. Wang, "Pattern classification of back-propagation algorithm using exclusive connecting network," *World Acad. Sci. Eng. Technol.*vol. 36, pp. 189–193, February, 2007.

[101] H. Najadat, S. Amani, and O. Ghadeer, "A new perfect hashing and pruning algorithm for mining association rule," *Bus. Transform. through Innov. Knowl. Manag. An Acad. Perspect. - Proc. 14th Int. Bus. Inf. Manag. Assoc. Conf. IBIMA*, vol. 4, no. January, pp. 2524–2531, 2010.

[102] F. Zhan, X. Zhu, L. Zhang, X. Wang, L. Wang, and C. Liu, "Summary of association rules," *IOP Conference Series: Earth and Environmental Science*, vol. 252, no. 3, p. 032219, 2019.

[103] W. Jentner and D. A. Keim, *High-Utility Pattern Mining*, Vol. 51, Springer International Publishing, , Berlin, 2019.

[104] P. Saraf, R. R Sedamkar, and S. Rathi, "PrefixSpan algorithm for finding sequential pattern with various constraints," *International Journal of Applied Information Systems*, vol. 9, no. 3, pp. 37–41, 2015.

[105] Y. Yuan and T. Huang, "A matrix algorithm for mining association rules," in *For Mining Association Rules*, pp. 370–379, Springer, Berlin, Heidelberg, 2005.

[106] D. Sirdeshmukh, N. B. Ahmad, M. S. Khan, and N. J. Ashill, "Drivers of user loyalty intention and commitment to a search engine: an exploratory study," *Journal of Retailing and Consumer Services*, vol. 44, pp. 71–81, 2018.

[107] J. Yadav and D. Kumar, "Subspace clustering using CLIQUE : an exploratory study," *Int. J. Adv. Res. Comput. Eng. Technol.*vol. 3, no. 2, pp. 372–378, 2014.

[108] N. M. Varghese and J. John, "Cluster optimization for enhanced web usage mining using fuzzy logic," in *Proceedings of the 2012 World Congr. Inf. Commun. Technol. WICT 2012*, pp. 948–952, Trivandrum, India, November 2012.

[109] K. Krishna and M. Narasimha Murty, "Genetic K-means algorithm," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433–439, 1999.

[110] M. Thelwall, "Quantitative comparisons of search engine results," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 11, pp. 1702–1710, 2008.

[111] E. Goldman, "Search engine bias and the demise of search engine utopianism," *Web Search 2003*, pp. 121–133, 2008.

[112] K. W.-T. Leung, W. Ng, and D. L. Dik Lun Lee, "Personalized concept-based clustering of search engine queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1505–1518, 2008.

[113] P. S. Sharma, D. Yadav, and P. Garg, "A systematic review on page ranking algorithms," *International Journal of Information Technology*, vol. 12, no. 2, pp. 329–337, 2020.