Review Article

Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey

Prem Sagar Sharma,1 Divakar Yadav,2 and R. N. Thakur3

1Uttarakhand Technical University, Dehradun, Uttarakhand, India
2National Institute of Technology, Hamirpur, Himachal Pradesh, India
3LBEF 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

Copyright © 2022 Prem Sagar Sharma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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...
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

<p>| Table 1: History of various search engines [2]. |</p>
<table>
<thead>
<tr>
<th>Search engine</th>
<th>Year</th>
<th>Methodology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gerard Salton [4]</td>
<td>1960s–1990s</td>
<td>—</td>
<td>Vector space model (VSM), inverse document frequency (IDF), term frequency (TF), term Discrimination values (TDV), and feedback mechanisms</td>
</tr>
<tr>
<td>Archie and Veronica [4]</td>
<td>1991–1992</td>
<td>—</td>
<td>&quot;Archie&quot; work for FTP sites &amp; &quot;Veronica&quot; 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</td>
</tr>
<tr>
<td>The first web directory [4]</td>
<td>1993–1994</td>
<td>—</td>
<td>ALLweb (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</td>
</tr>
<tr>
<td>Search directories [4]</td>
<td>1994–1995</td>
<td>It use large database which is created automatically</td>
<td>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</td>
</tr>
<tr>
<td>Yahoo [5]</td>
<td>1995</td>
<td>CORE (content optimization and relevance engine)</td>
<td>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</td>
</tr>
<tr>
<td>Meta-engines [4]</td>
<td>1995</td>
<td>It uses fusion to filter data The two main fusion methods: 1-collection fusion, 2-data fusion</td>
<td>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</td>
</tr>
<tr>
<td>Ask [4]</td>
<td>1996</td>
<td>It works on the basis of asking question instead of one or more words</td>
<td>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</td>
</tr>
<tr>
<td>Google [5]</td>
<td>1997</td>
<td>PageRank</td>
<td>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</td>
</tr>
<tr>
<td>Bing [4]</td>
<td>2009</td>
<td>Bing’s ranking algorithm utilizes machine learning/AI</td>
<td>MSN web portal developed in 2005 for in house search, but renamed in 2009 by Bing web search engine owned &amp; 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</td>
</tr>
</tbody>
</table>
pagerank 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 = Queryprocessor (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
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)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crawler</td>
<td>WebPages</td>
</tr>
<tr>
<td>WebPageRepository</td>
<td>Indexer</td>
</tr>
<tr>
<td>ContentAnalysis</td>
<td>IndexedWebPages</td>
</tr>
<tr>
<td>Meta-DataDescription</td>
<td>QueryProcessor</td>
</tr>
<tr>
<td>Query</td>
<td>RankedWebPages</td>
</tr>
</tbody>
</table>

Figure 2: General architecture of search engine.

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

\[ A \implies B, \quad (1) \]

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

For example, an association rule: Page A, Page B \( \implies \) 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.
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.

<table>
<thead>
<tr>
<th>WUM techniques</th>
<th>Algorithms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>K-means [8]</td>
<td>It is an unsupervised algorithm used for data mining and pattern recognition. The aim of K-Means is to minimize the cluster performance index</td>
</tr>
<tr>
<td></td>
<td>Greedy clustering using belief function [10, 11]</td>
<td>It is used to modelling evidence from expert opinions or statistical information</td>
</tr>
<tr>
<td></td>
<td>Improved fuzzy C mean [12]</td>
<td>It is a basic approach, used for image segmentation in which space divides into several clusters based on the pixel value of an image</td>
</tr>
<tr>
<td></td>
<td>CLIQUE (clustering in quest) [13]</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td>Cluster optimization using fuzzy cluster chase [14]</td>
<td>It is used to personalize web page clusters of end-users</td>
</tr>
<tr>
<td></td>
<td>K means with genetic algorithm—minimizes objective function [15]</td>
<td>The GKA is the most preferable algorithm for clustering to other evolutionary algorithms</td>
</tr>
<tr>
<td></td>
<td>Hierarchical agglomerative clustering [16]</td>
<td>It is a data exploratory analysis technique used in hierarchical clustering</td>
</tr>
<tr>
<td></td>
<td>Cluster optimization using anti-estimate approach [17]</td>
<td>It is used to remove redundant data that may occur during clustering</td>
</tr>
<tr>
<td></td>
<td>EB-DBSCAN (entropy-based DBSCAN) [18] DBSCAN [19]</td>
<td>It is used to identify the high-density regions/areas</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Web Page Repository</th>
<th>OUTPUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Analysis</td>
<td>Information from Link Mining</td>
</tr>
<tr>
<td></td>
<td>• Linked Based Classification</td>
</tr>
<tr>
<td></td>
<td>• Linked Based Cluster Analysis</td>
</tr>
<tr>
<td></td>
<td>• Link Type</td>
</tr>
<tr>
<td></td>
<td>• Link Strength</td>
</tr>
<tr>
<td></td>
<td>• Link Cardinality</td>
</tr>
<tr>
<td>Weight/Quality Calculator</td>
<td></td>
</tr>
</tbody>
</table>

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 (PUR) [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], Eigen rumor [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>
<thead>
<tr>
<th>Table 3: Summary of content mining algorithms.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content mining algorithms</strong></td>
</tr>
<tr>
<td>Decision trees</td>
</tr>
<tr>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>Support vector machine</td>
</tr>
<tr>
<td>Neural network</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: Summary of structure mining algorithms.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structure mining algorithm</strong></td>
</tr>
<tr>
<td>Page rank (PR)</td>
</tr>
<tr>
<td>Weighted page rank (WPR)</td>
</tr>
<tr>
<td>Eigen rumor (ER)</td>
</tr>
</tbody>
</table>
PageRank 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

\[
\text{PageRank}(A) = \sum_{B \in X_A} \left( \frac{\text{PageRank}(B)}{L(B)} \right).
\]  

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.
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 computed 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>
<thead>
<tr>
<th>Web mining techniques</th>
<th>Data</th>
<th>Algorithms</th>
<th>Methodology</th>
<th>List of input parameters</th>
<th>Complexity</th>
<th>Relevancy</th>
<th>Shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web structured mining (WSM) [53]</td>
<td>Hyperlinks, structure of documents</td>
<td>Page rank (PR)</td>
<td>Forward and backward links are used to compute the quality of the web pages</td>
<td>PR take O (log Time to compute the rank)</td>
<td>Return more relevant web pages</td>
<td>Return more relevant web pages as compared to the page rank algorithm</td>
<td>It is not considered the content of the web page</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weighted page rank (WPR)</td>
<td>Compute the weight of the pages based on their structure (links) and assign to the page. Finally, generate rank based on weight</td>
<td>The time complexity of WPR is O (log N)</td>
<td>It provides more relevant than WPR</td>
<td>Rely on web structure to compute rank</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Eigen rumor (ER)</td>
<td>It is the modified version of WPR by applying some other parameters</td>
<td>The time complexity of ER is log N</td>
<td>It return more relevant web pages</td>
<td>It is not considered the content of the web page</td>
<td></td>
</tr>
<tr>
<td>Web content mining (WCM) [53]</td>
<td>Text, image, audio, video, structure record</td>
<td>HITS</td>
<td>Monitor and consider those web pages which are visited by Internet users regularly</td>
<td>History of users' log files</td>
<td>The relevancy of this approach is moderate</td>
<td>Due to only rely on it is not much efficient</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weight links rank (WL rank)</td>
<td>It use the position of forward/backward links to compute the rank of web page</td>
<td>Link structure and content</td>
<td>—</td>
<td>—</td>
<td>It is not reliable because links are not placed at the proper location/position</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Distance ranking (DR)</td>
<td>The reinforcement learning algorithm is used by DR to compute the rank of the web page</td>
<td>After crawling, it computes the distance between 2-web pages</td>
<td>Time taken by DR is O (log N)</td>
<td>—</td>
<td>The logarithmic distance between 2-web pages is not reliable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tag rank (TR)</td>
<td>Hub &amp; authority concept are used to generate the rank</td>
<td>Forward, backward and tags on web pages</td>
<td>TR takes O (log N) time</td>
<td>Relevancy is moderate</td>
<td>It is less efficient</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dirichlet rank</td>
<td>It uses attributes of link like position, anchor text, tag where links exist, to compute the rank</td>
<td>Position of link, name of tags in which link exist, anchor text</td>
<td>—</td>
<td>—</td>
<td>The relevancy is not affected by the position of the link</td>
</tr>
<tr>
<td>Web usage mining (WUM) [53]</td>
<td>Web server logs, app server logs, app level logs</td>
<td>Query dependent ranking</td>
<td>It uses similar user queries to compute the rank</td>
<td>Training queries</td>
<td>Relevancy is high. It returns all the same web pages to a same user query</td>
<td>There are limited numbers of characteristics to find the similarity</td>
<td></td>
</tr>
<tr>
<td>S. No.</td>
<td>Year</td>
<td>Title</td>
<td>Journal</td>
<td>Author</td>
<td>Methodology/approach</td>
<td>Advantages</td>
<td>Limitations</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>2021</td>
<td>Using machine learning for web page classification in search Engine optimization</td>
<td>Future internet 2021, 13, 9</td>
<td>Goran Matosevi, Jasminka Dobsa, Dunja Mladeni</td>
<td>Authors used machine learning to classify web pages in SEO</td>
<td>Methods used in this research can help in building automated or semiautomated software for supporting SEO work</td>
<td>It is language-specific</td>
</tr>
<tr>
<td>2</td>
<td>2021</td>
<td>Learning to rank for educational search engines [29]</td>
<td>IEEE transactions</td>
<td>Arif Usta, Ismail Sengor Altingovde, Rıfat Ozcan, and Ozgur Ulusoy</td>
<td>Machine learning (ML) techniques. It is also called learning to Rank (LTR)</td>
<td>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</td>
<td>It has domain-specific features, and increases the perceived latency due to query dependence</td>
</tr>
<tr>
<td>3</td>
<td>2020</td>
<td>Big data analytics for search engine optimization [93]</td>
<td>MDPI, big data and cognitive computing</td>
<td>Ioannis C. Drivas, Damianos P. Sakas, Georgios A. Giannakopoulos, Daphne Kyriaki-Manessi</td>
<td>Author used agent based model, fuzzy cognitive mapping and big data analytics</td>
<td>It increases the organic search engine visits by using multiple SEO factors</td>
<td>This process of search engine optimization (SEO) could be a cost-effective</td>
</tr>
<tr>
<td>4</td>
<td>2020</td>
<td>Incremental refinement of page ranking of web pages [94]</td>
<td>International Journal of Information Retrieval Research, vol. 11, no. 2</td>
<td>P. S. Sharma, Divakar Yadav</td>
<td>Author used frequency of query keyword, hyperlink on query keywords and proxy server approach</td>
<td>It improves web page ranking and reduces perceived latency</td>
<td>Authors applied this approach only on web data</td>
</tr>
<tr>
<td>5</td>
<td>2019</td>
<td>Natural-language-based intelligent retrieval engine for BIM object database [95]</td>
<td>Computers in industry (Elsevier)</td>
<td>Songfei Wua, Qiyu Shena, Yichuan Deng, Jack Cheng</td>
<td>NLP (natural language processing)</td>
<td>It reduce threshold by using BIM object database</td>
<td>It is domain specific. Used ontology to understand semantic</td>
</tr>
<tr>
<td>6</td>
<td>2019</td>
<td>Jail-phish: An improved search engine-based phishing detection system [96]</td>
<td>Computers in industry (Elsevier)</td>
<td>Routhu Srinivasa Rao, Alwyn Roshan Pais</td>
<td>Author used heuristic technique to fetch similarity-based features to identify the phishing websites</td>
<td>It is used to identify the malicious users and also to detect phishing sites. It also works on free hosted websites</td>
<td>It is used as third-party based features to identify phishing sites on free hosted webs</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Forecasting tourist arrivals with machine learning and Internet search index [97]</td>
<td>Tourism management (Ekevier)</td>
<td>Shaolong Suna, Yunjie Weia, Kwok-Leung Tseu, Shouyang Wanga</td>
<td>Author used Machine learning (ML) and indexes of search engines</td>
<td>It increase forecasting accuracy and robustness</td>
<td>It is tested on 1- test case. It works on keyword selection</td>
</tr>
<tr>
<td>8</td>
<td>2019</td>
<td>An investigation of biases in web search engine query suggestions [98]</td>
<td>Online information review, Vol. 44 no. 2, 2020, pp. 365–381 © Emerald Publishing Limited</td>
<td>Malte Bonart, Anastasiasa Samokhina, Gernot Heisenberg and Philipp Schauer</td>
<td>Author designed a framework that automatically analyzes query suggestions for the web user</td>
<td>It is capable of Automatically collecting and analyzing query suggestions for a large repository of search keywords</td>
<td>It is topics derived. It is for the politician domain only</td>
</tr>
<tr>
<td>9</td>
<td>2018</td>
<td>Improving search engine optimization (SEO) by using hybrid Modified MCDM models [99]</td>
<td>Artificial intelligence review <a href="https://doi.org/10.1007/s10462-018-9644-0">https://doi.org/10.1007/s10462-018-9644-0</a> (Springer)</td>
<td>Hung-Jia Tsuei, Wei-Ho Tsai, Fu-Te Pan, Gwo-Hsiuhing Tseng</td>
<td>Multi-criteria decision-making (MCDM, also known as multicriteria decision analysis, MCDA)</td>
<td>Improving and evaluating search engine ranking</td>
<td>Need to improve on low-value websites</td>
</tr>
<tr>
<td>S. No</td>
<td>Year</td>
<td>Title</td>
<td>Journal</td>
<td>Author</td>
<td>Methodology/approach</td>
<td>Advantages</td>
<td>Limitations</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>---------</td>
<td>--------</td>
<td>-----------------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>10</td>
<td>2017</td>
<td>IBRI-CASONTO: Ontology-based semantic search engine [100]</td>
<td>Egyptian informatics journal (Elsevier)</td>
<td>Awny Sayed, Amal Al Muqrishi</td>
<td>Resource description framework (RDF) data &amp; ontological graph</td>
<td>It supports Arabic and English language. It use keyword-based search and a semantics-based search</td>
<td>It is domain specific. Indexing mechanism is not working for large data sets</td>
</tr>
<tr>
<td>11</td>
<td>2017</td>
<td>Death prediction and analysis using web mining techniques [101]</td>
<td>ICACCS - 2015, Coimbatore, India - 978-1-5090-459-4/17/31-EEE</td>
<td>Hesham Abdo, Ahmed Aqlan, et al.</td>
<td>Author used regression and neural network-based methodology to compute the rank of the webpages</td>
<td>It used the latest AI methodologies to improve page rank algorithms</td>
<td>Need to find suitable AI techniques to improve the prediction approach</td>
</tr>
<tr>
<td>12</td>
<td>2017</td>
<td>Design of a framework for knowledge based web page ranking [102]</td>
<td>International Journal of Engineering and Technology</td>
<td>P. Sharma, S. A.K, and P. Garg</td>
<td>Author used frequency of query keywords and proxy server approach</td>
<td>It reduce perceived latency</td>
<td>Synchronization problem between proxy and search engines</td>
</tr>
<tr>
<td>13</td>
<td>2016</td>
<td>A hybrid approach for extracting informative content from web pages [33]</td>
<td>Information processing and management (Elsevier)</td>
<td>Bahattin Vidinli, Rifat Ozcan</td>
<td>Authors designed a framework query suggestion</td>
<td>It can be reduced (simplified) to a problem of query compression</td>
<td>Need to be including spell checker to increase suggestion accuracy and needs to integrate content also</td>
</tr>
<tr>
<td>14</td>
<td>2015</td>
<td>Search-based QoS ranking prediction for web services in cloud environments [104]</td>
<td>Future generation computer systems (Elsevier)</td>
<td>Mao, Chengying Chen, Jifu Towey, Dave Chen, Jinfu Xie, Xiaoyuan</td>
<td>Author explore similarity measurement method for two ranked sequences. It predicts QoS ranking</td>
<td>The QoS information was used for ranking prediction</td>
<td>It does not use social relationships in the cloud platform which is important</td>
</tr>
<tr>
<td>15</td>
<td>2014</td>
<td>Effective ranking and search techniques for web resources considering semantic relationships [105]</td>
<td>Information processing and management (Elsevier)</td>
<td>Lee, Jihyun Min, Jun Ki Oh, Alice Chung, Chin Wan</td>
<td>Use ontology to compute weight for the semantic relationship</td>
<td>It increased the power of the query keyword for semantic relationship. It reduced the search space and increased the accuracy</td>
<td>It used ontology</td>
</tr>
<tr>
<td>16</td>
<td>2013</td>
<td>A hybrid approach for extracting informative content from web pages [33]</td>
<td>Information processing and management 49 (2013) 928–944 contents (Elsevier)</td>
<td>Erdinç Uzun, Hayri Volkan Agun, Tarik Yerlikaya</td>
<td>It uses decision tree learning to fetch informative contents and make rules. Extract reliable information by using these rules</td>
<td>It is very faster after making rules and provides high accuracy in results</td>
<td>It takes more time for rules creation the first time</td>
</tr>
<tr>
<td>17</td>
<td>2013</td>
<td>Topic-Driven SocialRank: Personalized search result ranking by identifying similar, credible users in a social network [35]</td>
<td>Knowledge-based systems 54 (2013) 230–242 contents (elsevier)</td>
<td>Young An Kim, Gun Woo Park</td>
<td>Focus on identifying similar users who have high credibility and sharing their search experiences</td>
<td>It is very useful to find more relevant search results by implicit hdp of familiar, credible users</td>
<td>It is tested on a small dataset</td>
</tr>
<tr>
<td>18</td>
<td>2012</td>
<td>Mining the real-time web: A novel approach to product recommendation [34]</td>
<td>Knowledge-based systems 29 (2012) 3–11 (Elsevier)</td>
<td>Garcia Espanza, Sandra O'Mahony, Michael P. Smyth, Barry</td>
<td>It uses a collaborative-filtering based approach</td>
<td>It is used for micro-blogging messages</td>
<td>It is not used by other domains like Twitter</td>
</tr>
<tr>
<td>19</td>
<td>2011</td>
<td>A music information system automatically generated via web content mining techniques Markus [36]</td>
<td>Information processing and management 47 (2011) 426–439 (Elsevier)</td>
<td>Markus Schedl, Gerhard Widmer, Peter knees, Tim Pohle</td>
<td>It uses web content mining techniques</td>
<td>It provides web-based access to a large collection of music artists. It is automated music information system.</td>
<td>It is domain specific (for music only)</td>
</tr>
<tr>
<td>20</td>
<td>2011</td>
<td>Snoogle: A search engine for pervasive environments [106]</td>
<td>IEEE transactions on parallel and distributed systems</td>
<td>Haodong Wang, Chiu C. Tan, and Qun Li Abstract—Embedding</td>
<td>It uses sensor networks, And communication overhead reduced by Bloom filters</td>
<td>In this, user can search a mobile object (s) that fit in detail</td>
<td>This system is not able to find a moving object in real-time</td>
</tr>
</tbody>
</table>
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 pre-fetching 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:

1. 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.
2. 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.
3. 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


