Web Page Ranking using Web Mining Techniques: A comprehensive survey

Prem Sagar Sharma¹, Divakar Yadav^{2*}

¹Uttarakhand Technical University, Dehradun (UK), India

²National Institute of Technology, Hamirpur (HP), India

premsagar1987@rediffmail.com, dsy99@rediffmail.com

ORCID ID: <https://orcid.org/0000-0001-6051-479X>

Abstract

Purpose: Due to the exponential growth of internet users and internet traffic, information seekers are highly dependent upon 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.

Design/methodology/approach: The search engine provides relevant information to internet users concerning to their query; based on content, link structure etc. However, it does not provide the guarantee of 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 analyse through Web structure mining (WSM) and their content, which analyses through Web content mining (WCM). Web mining plays an important role in computing the rank of web pages.

Findings: In this article, web mining types, techniques, tools, algorithms and their challenges are presented. Further, it provides a critical comprehensive survey for the researchers by presenting different features of web pages, which are important to check the quality of web pages.

Originality: In this work, authors presented different approaches/techniques, algorithms and evaluation approaches in previous researches and identified some critical issues in page ranking & web mining, which provide future directions for the researchers, working in the area.

Keywords: World Wide Web, Information retrieval, Web page ranking, Search engine, Web mining.

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 very difficult to analyze manually. Search engine[1][2][3] is an information retrieval tool for the Web like Google, Yahoo, Bing etc. The summary of various search engines is shown in table-1. Still, these search systems are sometimes not able to provide guarantee about reliable and accurate information but still, these systems provide better results than performing the task manually by experts. Many times these tools do not provide precise information because the IR system[4] returns information to internet users based on some specific retrieval criteria. For instance, it fetches web documents based on 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. Generally, a ranking mechanism creates the rank of web pages based on either keywords/reliability or links/popularity.

Hyperlinked Structure[5] has been developed in 1989 to share information among researchers in Switzerland. Later it becomes 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 as well as the people's lifestyles and economics of various countries.

The WWW is increasing exponentially (shown fig-1(a)) since its onset. 10 to 106 terabytes of traffic has increased in a month between 1995 & 2000. The total web traffic between the years 2005 & 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[6] of all internets was present (shown fig-1(b)). People not only view large amounts of video, but they also use high bandwidth due to viewing good quality videos.

All types of web content (like video, Netflix, webcam etc.) is generating demand. Now growing live videos is an important part of the internet. These video offerings from various sources like live Facebook, Twitter's broadcast, live YouTube, live sports etc. is expected to increase approximately 13% of traffic (shown fig-1(c)) of total video web traffic by 2021[6]. WWW is an important and widely used tool to provide reliable

information to internet user. It provides an important and easy mechanism for information (Like static text, images, dynamic & interactive services such as audio/video conferences etc.) sharing. It provides the facility to view various types of information including magazines, library resources in different sectors, current & business news. Now the web is an important source of all kinds of information.

But searching for information on the web is a difficult job for an information seeker. Web-based information retrieval systems called search engines [7], though 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.

Figure 1 (a). The exponential growth in terms of actual total internet Users [6]

Figure 1(b). Global Internet traffic from 2008-17

Figure 1(c). Predicted Global Internet traffic from 2018-21

It is important to understand that the term popularity is normally the result of link analysis and not user feedback. A web search engine (shown in figure-2) typically consists of a ranking system that measures the importance of Web Pages [8][9]. One can fetch content-based information from web documents using the hybrid approach [10]. The traffic of search engines is affected[11] by the following factors: Size of the web, loading speed (Page Redirect condition, Size of code) [12], Web security condition, SEO Crawling Factor (Title, heading, Meta Description of web page, Content, URL), User behaviour [8][13]. [14] presents a web page rank mechanism that is query dependent. This approach was much better and effective but it took more time to rank. In [15], 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 [16][17][18]. The main issue in content mining is that it was increasingly perceived latency, addressed this issue in [19] by an additional component said the proxy server.

Search Engines follow the following steps to process user queries:

- 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?

Figure 2. General Architecture of Search Engine

Working Process of Search Engine:

frontend_search_engine(UserQuery)

```
{
result_QP = Quesry_processor(UserQuery, Indexed_Web_Repository, Meta_data);
ranked_web_pages = Ranking_system(result_QP, Meta_data);
}
```

backend_search_engine (URL_List)

```
{
WebPageRepository = Crawler (URL_List);
indexed_web_page_Repository = indexer (webPageRepository);
new_list_of_URLs = contentAnalysis(webPageRepository);
Meta_data = contentAnalysis(webPageRepository);
Update_URL_List(URL_List, new_list_of_URLs);
}
```

QueryProcessor(UserQuery, Indexed_Web_Repository, Meta_data)

```
{
Query_Token= Tokenizing (UserQuery); // Tokenize user query as user input query
Parse_query = Parsing(Query_Token);
// user may use special operators in query , which are need to be remove
Stqk = stemming (Remove_Stop_keyword(Parse_query, thesaurus));
// remove stop keywords & do stemming
Precise_user_query = Query_expansion(Stqk);
}
```

```

// to search reliable information need to be convert a user query into technical query
list_of_webpage = Query_term_weighting(UserQuery, Indexed_Web_Repository, Meta_data)
return list_of_webpage;
}

```

Table 1. History of various search engines [20]

Search Engine	Year	Description
Gerard Salton	1960s - 1990s	Vector Space Model(VSM), Inverse Document Frequency (IDF), Term Frequency (TF), Term Discrimination Values(TDV), and Feedback Mechanisms
Archie and Veronica	1991-1992	'Archie' work for FTP sites & 'Veronica' work for Gopherspace. Gopherspace describes the aggregate of whole the information (like document file, papers, abstracts and other types of files) on the various Gopher servers in the world.
The First Web Directory	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	1994-1995	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	1995	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.
The Big-Guns	1995	If the "Yahoo" search engine fails to return a result then AltaVista search results automatically. It works for natural language queries and Boolean searches. AltaVista was big as well as fast also.
Meta-Engines	1995	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	1996	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	1997	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	2009	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.

2. WEB MINING

Data mining is used to find out relevant patterns or knowledge from repositories (like databases, texts, images etc.), which should be valid, useful and understandable. Text mining becomes popular and reliable by increasing the popularity of text documents. Web mining [21][22][23] 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 are used to collect the relevant information. Data mining techniques (shown in figure-3) [23][24]–[26] are used to fetch and discover relevant information automatically from web pages and web services in web mining. Data mining services are discussed in [27] to extract something useful out of the Web. There are following steps are needed to perform for this purpose:

- *Resource finding*: Extract the useful data/resources from either web documents, which are available online, or offline mode.
- *Information selection and pre-processing*: Apply the pre-processing (cleaning, normalization, feature extraction) on the specific information, which is automatically selected.
- *Transformation*: Pre-processed data is 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

- *Generalization*: It is used to fetch patterns from a website or across various websites by applying machine learning (ML) and other data mining techniques.
- *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 fig-4).

Figure 4. Architecture of Web Usages Mining (WUM)

The following techniques [28] are used in web usage mining:

2.1.1 Association Rules

Association rules in WUM provide the relationship between web pages that frequently appear next to one another in user sessions [29][23].

Statement of association rules written as:

$$A \Rightarrow B$$

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

For example, an association rule: Page A, Page B $\Rightarrow$ 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. The web usage mining process understands the existing data and behaviour of new instances. It identifies a particular class/category of a user. Classification techniques use Machine Learning (ML), Neural Network (NN) and statistical.

2.1.3 Clustering

Clustering techniques make sets of similar items from a large volumes of data by using distance functions which compute the similarity ratio between among items [29]. The contrast of the user/client and individual groups is a very important factor in such type of searching. There are two types of clustering, available in this area:

- User clustering
- Page clustering

User clustering is used to find those users which are having 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 found those patterns in which one set or set 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. There are some techniques that are utilized for sequential analysis [28] as shown in table-2. A detailed description of various algorithms of WUM Techniques are given in table-2.

Figure 5. Algorithms used in Web Usage Mining's Techniques

Table-2: Summary of Algorithms of WUM Techniques

WUM Techniques	Algorithms	Description
Clustering	K-Means [31]	It is an unsupervised algorithm used for data mining & pattern recognition, developed by J. B. MacQueen. The aim of K-Means is to minimize the cluster performance index.
	Greedy clustering using belief Function [32][33]	It is used to modelling evidence from expert opinions or statistical information.
	Improved Fuzzy C Mean [34]	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)[35]	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 [36]	It is used to personalize web page clusters of end-users.
	K means with genetic algorithm -Minimises objective function [37]	The GKA is the most preferable algorithm for clustering to other evolutionary algorithms.
	Hierarchical Agglomerative Clustering [38]	It is a data exploratory analysis technique used in Hierarchical clustering.
	Cluster Optimization using Ant-Nestmate Approach [39]	It is used to remove redundant data that may occur during clustering.
	EB-DBSCAN (Entropy-Based DBSCAN)[40]	It is used to identify the high-density regions/areas
	DBSCAN [41]	It is used to make clusters of arbitrary shapes.
Classification	Naïve Bayesia [42]	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 [43]	It is a classification technique used to construct decision trees for historical data.
	C4.5 [44]	It is a quick classification & high precision algorithm. It is used frequently for classification.
	SVM [45]	It is a classification algorithms that can be applied to linear and non-linear datasets.
	Backpropagation[46]	It is used as a gradient descent method to minimize error function in weight space.
Sequential	Hashing and pruning based algorithm [47]	It is a famous association rule mining technique to increase the performance of traditional Apriori algorithms.
	WAP tree association rule algorithm [48]	WAP Tree is a way to store the patterns in an effective manner by which these patterns are easily searchable.
	High Utility sequential patterns [49]	It is a data mining task that consists of a set of values having importance in a quantitative transaction database.
	PrefixSpan Algorithm [50]	It fetches sequential patterns using the pattern growth

		method. It works well for small datasets.
	Transaction Matrix comparison Algorithm[51]	It uses a boolean vector to discover frequent itemset. It required less memory because itemset stored in bits.

2.2 Web Content Mining (WCM)

Web Content Mining (shown in figure-6) (WCM)[52]–[54][10] is used to fetch relevant & Reliable information from web pages which may contain text documents, Hyperlinks, Structured data, audio & Video. Nowadays web pages are increasing exponentially over www.

Figure. 6 Architecture of Web Content Mining (WCM)

Fetching relevant data related to user queries from a large collection of web pages is very difficult and very time-consuming. Web content mining has the following approaches [52] to extract user relevant information from different types of data like unstructured data, structure data, Semi-structured etc. There are various content mining algorithms[53] used by the above content mining techniques are shown in table-2:

Table 2: Summary of Content Mining Algorithms

Content Mining Algorithms	Description
Decision Trees	It is a classification used by WCM & 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 a very powerful and easy to an understandable classifier
Support Vector Machine	It is a classification algorithms that can be applied to linear and non-linear 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 a input layer, hidden layers and an output layer. Each layer feeds to the next layer and the number of hidden layers are arbitrary.

2.3 Web Structure Mining (WSM)

Web Structure Mining (architecture shown in fig-7) detects the structural summary of a web page and its linked web pages. It finds out link (forward/backward) structure inside a web page by structure mining [52][55]. It is used to classify and compare web documents and integrate number of different web documents.

Web structure mining (WSM) (shown in figure-7) follow the following steps:

- Apply link analysis on a web page repository to extract links (forward/backward) summary of web pages.
- Apply a link mining techniques in the summary to find out the weight or quality of the web pages.

Figure 7. Architecture of Web Structure Mining (WSM)

Table-3: 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.

2.4 Challenges in Web Mining

Web mining is faced with some technical and non-technical issues. Non-technical issues occur due to management, fund and resources (such as professional humans), Some technical issues are discussed below:

- *Inappropriate data:* Collected data should be reliable and in proper format to do successfully mining because many times data is incomplete and unavailable. It is very difficult to assure the accuracy of such a data.
- *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.
- *Dynamic Web:* In dynamic web, data is frequently changed due to new updation. For example, sports data etc. Therefore, the complexity of mining is increased.
- *Shortage of Mining Tools:* Need to develop a mining tools because a very less 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 [11][56][57]–[60] of the search engines (shown in table-4). 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[52]. It has main three features:

- Frequency of query keywords inside title
- Frequency of query keywords inside heading tags (H1, H2.....H6) separately.
- 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 like headings, links/URLs etc.

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

Table 4: 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.<="" td=""></h6>>
		Freq_QK_paragraph	Number of query keywords in paragraph <p> </p> 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

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 [61]. The search engines have become the driver of internet users that move them towards the highly ranked web by using various web mining techniques[62]. 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 WPR (Web Pages Ranking) algorithms is explained in [63]. There are a several ways to improve the ranking of a web page on search engines, as SPAM farms are a very famous method to enhance a Web site's ranking. During Rank calculation of web pages, cognitive spammer framework (CSF) delete all spam web documents [64]. A framework Preference-based Universal Ranking Integration (PURI) [65] is designed by combining various ranking mechanisms. The internet is an important source to access information from the web. While almost all web pages contain a lot of noise in web like: advertisements, different types of banners, unreliable links etc. that affect the performance of content and structure based search engines, Question -Answering System, Web Summarization [10]. For instance, it

fetches web documents based on 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 [66]. Ranking system utilize various web page ranking algorithms like page rank [15][67], weighted page rank [68], Eigenrumor [69], HITS [70], Weight Links Rank [18], distance ranking [71], tag rank [72], query dependent[14] to compute a rank of web page. It returns the order of web pages (order is done based on their rank). Table 5 shows the summary of web mining techniques and ranking algorithms for each mining technique.

Table-5: 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)[73]	Hyperlinks, Structure of Documents	Page Rank (PR)	Forward & backward Links are used to compute the quality of the web pages.	Forward and backward links	PR take $O(\log N)$ 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)[73]	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 & Content	*	*	It is not reliable because links are not placed at the proper location/ position.
		Distance Ranking(DR)	The reinforcement	After crawling, it	Time taken by DR is	*	The logarithmic

			learning algorithm is used by DR to compute the rank of the web page.	compute the distance between 2-web pages.	$O(\log N)$		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)[73]	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.
		SimRank	Design a rank model for the same type of user queries	Training queries	*	Relevancy is high. it returns all the same web pages to the same user query	There are limited numbers of characteristics to find the similarity.

4. LITERATURE REVIEW

Due to increasing the information for humans on the WWW, the responsibility of the internet also increased. It is very easy 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 very difficult 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 to users to select reliable information.

In [74], authors give an approach to fetch experts attributes by using text mining from the web i.e. it is a recommended model to return a precise record. This research has shown the effectiveness of the proposed approach is box-office revenue prediction. In [75] 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 [76] the author developed a framework for Geographic Information Mining (GIM) framework. Microsoft discussion(MSD) forums used ExpertRank [77], 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 [78] author identified user activities in the SO- forum and compared with their GitHub repositories, and feasible features of user (active in both platforms). In [79] author proposed user activity models for stackoverflow, Wenwo Forums & SinaWeibo to classify real experts. In [80] 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. 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[81], to find experts, citation networks [82] & email communications[83]. For online users, in [84] 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 & HITS etc. algorithms were used to compute social network authority scores. A search engine to fetch biomedical information [58] 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. Table 6 presents the summary of various research papers based on different attributes like methodologies, approaches, pros and cons etc.

Table 6 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 [85]	Future Internet 2021, 13, 9	Goran Matosevi, Jasminka Dobsa, Dunja Mladen	Authors used machine learning to classify web pages in SEO.	Methods used in this research can help in building automated or semi-automated software for supporting SEO work.	It is language-specific.
2	2021	Learning to Rank for Educational Search Engines [11]	IEEE Transactions	Arif Usta, Ismail Sengor Altinogvde, Rifat Ozcan, and Ozgur Ulusoy	Machine Learning (ML) techniques. It is also called learning to rank (LTR).	It expose 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 [86]	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[19]	Int. J. Inf. Retr. Res., 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 webpage ranking and reduces perceived latency.	Authors applied this approach only on web data.
5	2019	Natural-language-based intelligent retrieval	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

		engine for BIM object database [87]					semantic.
6	2019	Jail-Phish: An improved search engine based phishing detection system [88]	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 [89]	Tourism Management (Elsevier)	Shaolong Suna, Yunjie Weia, Kwok-Leung Tsuc, 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 [90]	Online Information Review, Vol. 44 No. 2, 2020, pp. 365-381 © Emerald Publishing Limited	Malte Bonart, Anastasia Samokhina, Ger not 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 MCDM models [8]	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 Multi-Criteria Decision Analysis, MCDA).	Improving and evaluating search engine ranking.	Need to improve on low-value websites.
10	2017	IBRI-CASONT O: Ontology-based semantic search engine [91]	Egyptian Informatics Journal (Elsevier)	Awany 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	ICACCS - 2015, Coimbatore,	Hesham Abdo, Ahmed Aqlan, Et. al.	Author used regression and Neural Network based	It used the latest AI methodologies to improve page	Need to find suitable AI techniques to

		Analysis Using Web Mining Techniques [92]	INDIA- 978-1-5090-4559-4/17/\$31-EEE		methodology to compute the rank of the webpages.	rank algorithms.	improve the prediction approach.
12	2017	Design of a Framework for Knowledge Based Web Page Ranking [93]	<i>Int. J. Eng. Technol</i>	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[94]	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[95]	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 [96]	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[10]	Information Processing and Management 49 (2013) 928–944 Contents (Elsevier)	Erdoğan 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	Knowledge-Based Systems 54 (2013) 230–242 Contents	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	It is tested on a small dataset.

		result ranking by identifying similar, credible users in a social network [97]	(Elsevier)			users.	
18	2012	Mining the real-time web: A novel approach to product recommendation[98]	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 [99]	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[100]	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.

4.1 Some observations have been seen in available literature reviews, which are discussed below:

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 prefetching 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, i.e. 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 require approaches that personalize the content of web pages with respect to the user's group.

5. CONCLUSION AND FUTURE SCOPE

Three categories of ranking algorithms are mainly discussed. The first category of algorithm, which is based on content of web pages is known as content-based page ranking. The second category of algorithm, which uses link structure of world wide web is known as web structure-based page ranking algorithms and the third category used 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 some other challenges in classification and clustering techniques.

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

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

- Web structure-based page ranking algorithms may ignore web pages, which have less page ranking score but have 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.
- The size of WWW is very large 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.

References

- [1] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important factors for improving Google search rank," *Futur. Internet*, vol. 11, no. 2, 2019, doi: 10.3390/fi11020032.
- [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," *Futur. Gener. Comput. Syst.*, vol. 92, pp. 972–986, 2019, doi: 10.1016/j.future.2018.01.037.
- [3] D. Qiao, J. Zhang, Q. Wei, and G. Chen, "Finding competitive keywords from query logs to enhance search engine advertising," *Inf. Manag.*, vol. 54, no. 4, pp. 531–543, 2017, doi: 10.1016/j.im.2016.11.003.
- [4] Z. Hadjilambrou, M. Kleanthous, G. Antoniou, A. Portero, and Y. Sazeides, "Comprehensive Characterization of an Open Source Document Search Engine," *ACM Trans. Archit. Code Optim.*, vol. 16, no. 2, 2019, doi: 10.1145/3320346.
- [5] K. Mohan, "A Survey on Web Structure Mining," *Int. J. Adv. Comput. Res.*, vol. 1, no. 1, pp. 715–720, 2017.
- [6] "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.
- [7] N. Höchstötter and D. Lewandowski, "What users see - Structures in search engine results pages," *Inf. Sci. (Ny)*, vol. 179, no. 12, pp. 1796–1812, 2009, doi: 10.1016/j.ins.2009.01.028.
- [8] H. J. Tsuei, W. H. Tsai, F. Te Pan, and G. H. Tzeng, "Improving search engine optimization (SEO) by using hybrid modified MCDM models," *Artif. Intell. Rev.*, vol. 53, no. 1, 2020, doi: 10.1007/s10462-018-9644-0.
- [9] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Inf. Process. Manag.*, vol. 42, no. 1 SPEC. ISS, pp. 248–263, 2006, doi: 10.1016/j.ipm.2004.10.007.
- [10] E. Uzun, H. V. Agun, and T. Yerlikaya, "A hybrid approach for extracting informative content from web pages," *Inf. Process. Manag.*, vol. 49, no. 4, pp. 928–944, 2013, doi: 10.1016/j.ipm.2013.02.005.
- [11] A. Usta, I. S. Altıngöç, R. Özcan, and O. Ulusoy, "Learning to Rank for Educational Search Engines," *IEEE Trans. Learn. Technol.*, vol. 0050, no. c, pp. 1–14, 2021, doi: 10.1109/TLT.2021.3075196.
- [12] O. Dan and B. D. Davison, "Measuring and predicting search engine users' satisfaction," *ACM Comput. Surv.*, vol. 49, no. 1, 2016, doi: 10.1145/2893486.
- [13] 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 Inf. Rev.*, vol. 40, no. 2, pp. 239–255, 2016, doi: 10.1108/OIR-04-2015-0112.
- [14] L. W. Lee, J. Y. Jiang, C. Der Wu, and S. J. Lee, "A query-dependent ranking approach for search engines," *2nd Int. Work. Comput. Sci. Eng. WCSE 2009*, vol. 1, pp. 259–263, 2009, doi: 10.1109/WCSE.2009.666.
- [15] R. Baeza-Yates and E. Davis, "Web Page Ranking using Link Attributes Categories and Subject Descriptors," *Proc. 13th Int. World Wide Web Conf. Altern. track Pap. posters*, pp. 328–329, 2004.
- [16] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G. V. Uma, and K. Sarukesi, "Relevance ranking and evaluation of search results through web content mining," *Lect. Notes Eng. Comput. Sci.*, vol. 2195, pp. 456–460, 2012.

- [17] R. K. Roul, S. R. Asthana, and G. Kumar, "Spam web page detection using combined content and link features," *Int. J. Data Mining, Model. Manag.*, vol. 8, no. 3, pp. 209–222, 2016, doi: 10.1504/IJDMMM.2016.079063.
- [18] P. Sharma, D. Tyagi, and P. Bhadana, "Weighted Page Content Rank for Ordering Web Search Result," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 12, pp. 7301–7310, 2010.
- [19] P. S. Sharma and D. Yadav, "Incremental Refinement of Page Ranking of Web Pages," *Int. J. Inf. Retr. Res.*, vol. 10, no. 3, pp. 57–73, 2020, doi: 10.4018/ijrr.2020070104.
- [20] T. Seymour, D. Frantsvog, and S. Kumar, "History Of Search Engines," *Int. J. Manag. Inf. Syst.*, vol. 15, no. 4, p. 47, 2011, doi: 10.19030/ijmis.v15i4.5799.
- [21] D. Bollegala, Y. Matsuo, and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 977–990, 2011, doi: 10.1109/TKDE.2010.172.
- [22] 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," *J. Web Semant.*, vol. 9, no. 4, pp. 365–401, 2011, doi: 10.1016/j.websem.2011.06.004.
- [23] M. Jawad and H. Mughal, "Data Mining : Web Data Mining Techniques , Tools and Algorithms : An Overview," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 208–215, 2018.
- [24] 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, doi: 10.1016/j.cmpb.2007.01.007.
- [25] A. Anagnostopoulos, A. Z. Broder, and D. Carmel, "Sampling search-engine results," *World Wide Web*, vol. 9, no. 4, pp. 397–429, 2006, doi: 10.1007/s11280-006-0222-z.
- [26] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Comput. Networks*, vol. 50, no. 10, pp. 1448–1463, 2006, doi: 10.1016/j.comnet.2005.10.020.
- [27] M. Zorrilla and D. García-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners," *Decis. Support Syst.*, vol. 55, no. 1, pp. 399–411, 2013, doi: 10.1016/j.dss.2012.05.045.
- [28] P. Suthar and P. B. Oza, "A Survey on Web Usage Mining Techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 10, pp. 3824–3829, 2015.
- [29] M. Kumari and S. Soni, "A Review of classification in Web Usage Mining using K- Nearest Neighbour," *Adv. Comput. Sci. Technol.*, vol. 10, no. 5, pp. 1405–1416, 2017.
- [30] C. Vyas, "Evaluating state tourism websites using Search Engine Optimization tools," *Tour. Manag.*, vol. 73, no. January, pp. 64–70, 2019, doi: 10.1016/j.tourman.2019.01.019.
- [31] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012, doi: 10.1016/j.phpro.2012.03.206.
- [32] L. C. Principle, "Methods for building belief functions," pp. 1–67, 2011.
- [33] 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, doi: 10.1089/big.2020.0133.
- [34] H. Rustum, H. Hadi, and A. AbdulZahraa, "Improved Fuzzy C-Mean Algorithm for Image Segmentation," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 6, 2016, doi: 10.14569/ijarai.2016.050602.
- [35] 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.
- [36] N. M. Varghese and J. John, "Cluster optimization for enhanced web usage mining using fuzzy logic," *Proc. 2012 World Congr. Inf. Commun. Technol. WICT 2012*, pp. 948–952, 2012, doi: 10.1109/WICT.2012.6409211.
- [37] K. Krishna and M. N. Murty, "Genetic K-means algorithm," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 29, no. 3, pp. 433–439, 1999, doi: 10.1109/3477.764879.
- [38] P. Contreras and F. Murtagh, "Hierarchical clustering," *Handb. Clust. Anal.*, no. February, pp. 103–124, 2015, doi: 10.1201/b19706.
- [39] A. Alphy and S. Prabakaran, "Cluster optimization for improved web usage mining using ant nestmate approach," *Int. Conf. Recent Trends Inf. Technol. ICRTIT 2011*, pp. 1271–1276, 2011, doi: 10.1109/ICRTIT.2011.5972369.
- [40] T. Wang, C. Ren, Y. Luo, and J. Tian, "NS-DBSCAN: A density-based clustering algorithm in network space," *ISPRS Int. J. Geo-Information*, vol. 8, no. 5, 2019, doi: 10.3390/ijgi8050218.
- [41] M. Daszykowski and B. Walczak, "Density-Based Clustering Methods," *Compr. Chemom.*, vol. 2, pp. 635–654, 2009, doi: 10.1016/B978-044452701-1.00067-3.
- [42] D. Berrar, "Bayes' theorem and naive bayes classifier," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.
- [43] D. Steinberg, "CART : Classification and Regression Trees," no. January 2009, 2015.
- [44] I. I. Sinam and A. Lawan, "An improved C4.5 model classification algorithm based on Taylor's series,"

- Jordanian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 34–42, 2019, doi: 10.5455/jjcit.71-1546551963.
- [45] M. Awad and R. Khanna, “Efficient learning machines: Theories, concepts, and applications for engineers and system designers,” *Effic. Learn. Mach. Theor. Concepts, Appl. Eng. Syst. Des.*, no. January, pp. 1–248, 2015, doi: 10.1007/978-1-4302-5990-9.
 - [46] I. Jung and G. Wang, “Pattern Classification of Back-Propagation Algorithm Using Exclusive Connecting Network,” *World Acad. Sci. Eng. Technol.*, vol. 36, no. February, pp. 189–193, 2007.
 - [47] 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 2010*, vol. 4, no. January, pp. 2524–2531, 2010, doi: 10.5171/2011.6512178.
 - [48] F. Zhan, X. Zhu, L. Zhang, X. Wang, L. Wang, and C. Liu, “Summary of Association Rules,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 252, no. 3, 2019, doi: 10.1088/1755-1315/252/3/032219.
 - [49] W. Jentner and D. A. Keim, *High-Utility Pattern Mining*, vol. 51, no. May. Springer International Publishing, 2019.
 - [50] P. Saraf, R. R Sedamkar, and S. Rathi, “PrefixSpan Algorithm for Finding Sequential Pattern with Various Constraints,” *Int. J. Appl. Inf. Syst.*, vol. 9, no. 3, pp. 37–41, 2015, doi: 10.5120/ijais15-451380.
 - [51] Y. Yuan and T. Huang, “for Mining Association Rules,” pp. 370–379, 2005.
 - [52] F. Johnson, “Web Content Mining Techniques : A Survey,” *Int. J. Comput. Appl.*, vol. 47, no. 11, pp. 44–50, 2012.
 - [53] 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, [Online]. Available: <http://www.ijctjournal.org>.
 - [54] A. kumar, “A Study on Web Content Mining,” *Int. J. Eng. Comput. Sci.*, vol. 6, no. 1, pp. 2015–2018, 2017, doi: 10.18535/ijecs/v6i1.29.
 - [55] 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,” *Comput. Human Behav.*, vol. 78, pp. 223–234, 2018, doi: 10.1016/j.chb.2017.10.004.
 - [56] 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,” *J. Retail. Consum. Serv.*, vol. 44, no. June, pp. 71–81, 2018, doi: 10.1016/j.jretconser.2018.06.002.
 - [57] M. Thelwall, “Quantitative comparisons of search engine results,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 11, pp. 1702–1710, 2008, doi: 10.1002/asi.20834.
 - [58] 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, doi: 10.1093/bioinformatics/btn469.
 - [59] E. Goldman, “Search Engine Bias and the Demise of Search Engine Utopianism,” no. Sullivan 2003, pp. 121–133, 2008, doi: 10.1007/978-3-540-75829-7_8.
 - [60] K. W. T. Leung, W. Ng, and D. L. Lee, “Personalized concept-based clustering of search engine queries,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 11, pp. 1505–1518, 2008, doi: 10.1109/TKDE.2008.84.
 - [61] C. MacDonald and I. Ounis, “The influence of the document ranking in expert search,” *Inf. Process. Manag.*, vol. 47, no. 3, pp. 376–390, 2011, doi: 10.1016/j.ipm.2010.09.004.
 - [62] K. C. Lee and S. Lee, “Interpreting the web-mining results by cognitive map and association rule approach,” *Inf. Process. Manag.*, vol. 47, no. 4, pp. 482–490, 2011, doi: 10.1016/j.ipm.2010.11.005.
 - [63] V. Derhami, E. Khodadadian, M. Ghasemzadeh, and A. M. Zareh Bidoki, “Applying reinforcement learning for web pages ranking algorithms,” *Appl. Soft Comput. J.*, vol. 13, no. 4, pp. 1686–1692, 2013, doi: 10.1016/j.asoc.2012.12.023.
 - [64] A. Makkar and N. Kumar, “An efficient deep learning-based scheme for web spam detection in IoT environment,” *Futur. Gener. Comput. Syst.*, vol. 108, pp. 467–487, 2020, doi: 10.1016/j.future.2020.03.004.
 - [65] 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 Syst.*, vol. 49, pp. 22–36, 2013, doi: 10.1016/j.knosys.2013.04.007.
 - [66] M. S. Faisal, A. Daud, A. U. Akram, R. A. Abbasi, N. R. Aljohani, and I. Mehmood, “Expert ranking techniques for online rated forums,” *Comput. Human Behav.*, vol. 100, no. December 2017, pp. 168–176, 2019, doi: 10.1016/j.chb.2018.06.013.
 - [67] T. W. L. Page, S. Brin, R. Motavni, “The PageRank Citation Ranking: Bringing Order to the Web,” *Tech. Report, Stanford Digit. Libr. Technol.*, 1998, doi: 10.1109/IISWC.2012.6402911.
 - [68] W. Xing and A. Ghorbani, “Weighted PageRank algorithm,” *Proc. - Second Annu. Conf. Commun. Networks Serv. Res.*, pp. 305–314, 2004.

- [69] K. Fujimura, T. Inoue, and M. Sugisaki, "The EigenRumor Algorithm for Ranking Blogs," *Proc. Second Annu. Work. Weblogging Ecosyst. Aggregation, Anal. Dyn. Chiba, Japan*, pp. 59–74, 2005.
- [70] and J. K. S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, "Mining the Link Structure of the World Wide Web," *Computer (Long. Beach. Calif.)*, vol. 12, no. 23, pp. 60–67, 1999, doi: 10.1300/j101v12n23_04.
- [71] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages," *Inf. Process. Manag.*, vol. 44, no. 2, pp. 877–892, 2008, doi: 10.1016/j.ipm.2007.06.004.
- [72] ShenJie, ChenChen, Zhanghui, S. Rong-Shuang, ZhuYan, and HeKun, "Tag Rank: A new rank algorithm for webpage based on social web," *Proc. Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2008*, pp. 254–258, 2008, doi: 10.1109/ICCSIT.2008.45.
- [73] P. S. Sharma, D. Yadav, and P. Garg, "A systematic review on page ranking algorithms," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 329–337, 2020, doi: 10.1007/s41870-020-00439-3.
- [74] 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," *Futur. Gener. Comput. Syst.*, vol. 111, pp. 859–872, 2020, doi: 10.1016/j.future.2019.08.033.
- [75] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews," *Inf. Process. Manag.*, vol. 57, no. 5, p. 102278, 2020, doi: 10.1016/j.ipm.2020.102278.
- [76] Y. Zhang *et al.*, "Extracting geographic features from the Internet: A geographic information mining framework," *Knowledge-Based Syst.*, vol. 174, pp. 57–72, 2019, doi: 10.1016/j.knosys.2019.02.031.
- [77] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "ExpertRank : A topic-aware expert finding algorithm for online knowledge communities," *Decis. Support Syst.*, vol. 54, no. 3, pp. 1442–1451, 2013, doi: 10.1016/j.dss.2012.12.020.
- [78] B. Vasilescu, V. Filkov, U. C. Davis, and A. Serebrenik, "StackOverflow and GitHub : Associations Between Software Development and Crowdsourced Knowledge," 2013, doi: 10.1109/SocialCom.2013.35.
- [79] L. Zhe and B. J. Jansen, "Questioner or question : Predicting the response rate in social question and answering on Sina Weibo," vol. 54, no. September 2017, pp. 159–174, 2018, doi: 10.1016/j.ipm.2017.10.004.
- [80] D. Madrid, I. Cantador, A. Bellogín, M. E. Cortés-cediel, and O. Gil, "Personalized recommendations in e-participation : Offline experiments for the Personalized recommendations in e-participation : Offline experiments for the ' Decide Madrid ' platform," no. August, 2017, doi: 10.1145/3127325.3127330.
- [81] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert Finding for Question Answering via Graph Regularized Matrix Completion," vol. 4347, no. c, pp. 1–12, 2014, doi: 10.1109/TKDE.2014.2356461.
- [82] R. N. Srivastava, Divya., and Mall, "Structural Analysis of L-Bracket using ANSYS," *i-manager's J. Mech. Eng.*, vol. 7, no. 2, 2017, doi: <https://doi.org/10.26634/jme.7.2.13432>.
- [83] H. Kareem and L. Asker, "Detecting Hierarchical Ties Using Link-Analysis Ranking at Different Levels of Time Granularity," pp. 1–4.
- [84] 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*, 2007, pp. 111–114.
- [85] G. Matošević, J. Dobša, and D. Mladenčić, "Using machine learning for web page classification in search engine optimization," *Futur. Internet*, vol. 13, no. 1, pp. 1–20, 2021, doi: 10.3390/fi13010009.
- [86] I. C. Drivas, D. P. Sakas, G. A. Giannakopoulos, and D. Kyriaki-Manessi, "Big data analytics for search engine optimization," *Big Data Cogn. Comput.*, vol. 4, no. 2, pp. 1–22, 2020, doi: 10.3390/bdcc4020005.
- [87] S. Wu, Q. Shen, Y. Deng, and J. Cheng, "Natural-language-based intelligent retrieval engine for BIM object database," *Comput. Ind.*, vol. 108, pp. 73–88, 2019, doi: 10.1016/j.compind.2019.02.016.
- [88] R. S. Rao and A. R. Pais, "Jail-Phish: An improved search engine based phishing detection system," *Comput. Secur.*, vol. 83, pp. 246–267, 2019, doi: 10.1016/j.cose.2019.02.011.
- [89] S. Sun, Y. Wei, K. L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," *Tour. Manag.*, vol. 70, no. July 2018, pp. 1–10, 2019, doi: 10.1016/j.tourman.2018.07.010.
- [90] M. Bonart, A. Samokhina, G. Heisenberg, and P. Schaer, "An investigation of biases in web search engine query suggestions," *Online Inf. Rev.*, vol. 44, no. 2, pp. 365–381, 2020, doi: 10.1108/OIR-11-2018-0341.
- [91] A. Sayed and A. Al Muqrishi, "IBRI-CASONTO: Ontology-based semantic search engine," *Egypt. Informatics J.*, vol. 18, no. 3, pp. 181–192, 2017, doi: 10.1016/j.eij.2017.01.001.
- [92] A. D. Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, "Death Prediction and Analysis Using Web Mining Techniques," 2017.

- [93] P. Sharma, S. A.K, and P. Garg, "Design of a Framework for Knowledge Based Web Page Ranking," *Int. J. Eng. Technol.*, vol. 9, no. 3, pp. 2236–2244, 2017, doi: 10.21817/ijet/2017/v9i3/1709030219.
- [94] I. B. Vidinli and R. Ozcan, "New query suggestion framework and algorithms: A case study for an educational search engine," *Inf. Process. Manag.*, vol. 52, no. 5, pp. 733–752, 2016, doi: 10.1016/j.ipm.2016.02.001.
- [95] C. Mao, J. Chen, D. Towey, J. Chen, and X. Xie, "Search-based QoS ranking prediction for web services in cloud environments," *Futur. Gener. Comput. Syst.*, vol. 50, pp. 111–126, 2015, doi: 10.1016/j.future.2015.01.008.
- [96] J. Lee, J. K. Min, A. Oh, and C. W. Chung, "Effective ranking and search techniques for Web resources considering semantic relationships," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 132–155, 2014, doi: 10.1016/j.ipm.2013.08.007.
- [97] 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 Syst.*, vol. 54, pp. 230–242, 2013, doi: 10.1016/j.knosys.2013.09.011.
- [98] S. Garcia Esparza, M. P. O'Mahony, and B. Smyth, "Mining the real-time web: A novel approach to product recommendation," *Knowledge-Based Syst.*, vol. 29, pp. 3–11, 2012, doi: 10.1016/j.knosys.2011.07.007.
- [99] M. Schedl, G. Widmer, P. Knees, and T. Pohle, "A music information system automatically generated via Web content mining techniques," *Inf. Process. Manag.*, vol. 47, no. 3, pp. 426–439, 2011, doi: 10.1016/j.ipm.2010.09.002.
- [100] H. Wang, C. C. Tan, and Q. Li, "Snoogle: A search engine for pervasive environments," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 8, pp. 1188–1202, 2010, doi: 10.1109/TPDS.2009.145.