ISSN: 0067-2904

Links Evaluation and Ranking Based on Semantic Metadata Analysis

Matheel E. Abdulmunim, Esraa Q. Naamha*
Department of Computer Science, University of technology- Iraq, Baghdad, Iraq

Received: 20/11/2022 Accepted: 6/5/2023 Published: 30/5/2024

Abstract

There is a vast quantity of information contained within the billions of web pages that make up the World Wide Web (WWW). Search engines carry out a variety of activities depending on their own architectures for retrieving the necessary information from the WWW. The search engine typically returns a huge number of pages in response to a user's query when the user submits one. Numerous ranking techniques have been utilized on search results to aid consumers in navigating the result list. The majority of ranking algorithms described in literature are either content-based or link-based, and they do not take user usage patterns into account. The presented study discusses web mining ranking algorithms depending on structures, contents, and usages and suggests a new ranking method to assess the significance of links with the use of semantic metadata analysis, which considers the number of links visited across nearly all regions, time periods, and related topics and queries. Furthermore, the suggested system uses the user's query to find more relevant information. The most valuable pages are thus displayed at the top of the result list based on user browsing behavior, which significantly decreases the search space. Results showed better ranking output based on different criteria, such as the number of links visited yearly, the number of links visited hourly, the number of links visited by region, the number of links visited by related topics, and the number of links visited by related queries.

Keywords: Programmable (CSE), JSON API, PageRank, Web Mining, Information retrieval, Semantic Metadata.

تقييم الروابط وترتيبها بناءً على تحليل البيانات الوصفية الدلالية

مثيل عماد الدين عبد المنعم، اسراء قاسم نعمة*
قسم علوم الحاسوب، الجامعة التكنولوجية– العراق، بغداد، العراق.

الخلاصة

تتكون شبكة الوب العالمية من بلايين من صفحات الوب وتحتوي على قدر هائل من المعلومات المتاحة داخل صفحات الوب. لاسترداد المعلومات المطلوبة من شبكة الوب العالمية ، تؤدي محركات البحث عددًا من المهام بناءً على هياكلها الخاصة. عندما يرسل المستخدم استعلامًا إلى محرك البحث ، فإنه يعرض بشكل عام عددًا كبيرًا من الصفحات استجابةً لطلب بحث المستخدم. لدعم تنقل المستخدمين في قائمة النتائج ، يتم تطبيق طرق ترتيب مختلفة على نتائج البحث. معظم خوارزميات الترتيب الواردة في الأدبيات إما قائمة على الارتباط أو قائمة على المحتوى ولا تأخذ في الاعتبار اتجاهات استعمال المستخدم. يناقش هذا البحث خوارزميات الترتيب بناءً على المحتويات والهياكل والاستعمالات في التنقيب على الويب وتقترح نهجًا جديدًا للترتيب لتقييم

*Email: cs.20.29@grad. uotechnology.edu.iq

أهمية الروابط باستعمال تحليل البيانات الوصفية الدلالية ، والذي يأخذ في الاعتبار عدد الروابط التي تمت زيارتها عبر جميع الفترات الزمنية والمناطق و الموضوعات والاستفسارات ذات الصلة. يستعمل النظام المقترح للعثور على المزيد من المعلومات ذات الصلة وفقًا لاستعلام المستخدم. لذا ، فإن هذا المفهوم مفيد جدًا لعرض الصفحات الأكثر قيمة في أعلى قائمة النتائج على أساس سلوك تصفح المستخدم ، مما يقلل من مساحة البحث على نطاق واسع. أظهرت النتائج ترتيبًا أفضل للمخرجات بناءً على معايير مختلفة ، مثل عدد الروابط التي تمت زيارتها سنويًا ، وعدد الروابط التي تمت زيارتها كل ساعة ، وعدد الروابط التي تمت زيارتها حسب المنطقة ، وعدد الروابط التي تمت زيارتها حسب الموضوعات ذات الصلة ، وعدد الروابط التي تمت زيارتها بواسطة استفسارات ذات صلة.

1. Introduction

There is an enormous quantity of information available on the billions of web pages that make up the WWW. Depending on their designs, search engines carry out some operations to retrieve the needed information from the Web [1]. These procedures could be challenging and time-consuming. The steps taken by each search engine include information crawling, searching, indexing, and sorting or ranking. A crawler browses and downloads all the webpages of a website in order to get the necessary data [2] and [3]. The information produced by the crawler must be saved in a certain way in order for the search engine to retrieve it; the information is indexed to cut down on the amount of time required to look at it. The user interface that is required to allow the user to query information has been represented by the web search engine. It is the channel through which the information repository and a user are connected [4] [5]. There are a huge number of web pages that are relevant to a user's search query when they submit it to a search engine. However, the user only requires a few web pages to function properly. Even so, this number is enormous (in millions). Before the results are shown, a search engine sorts the results using a ranking algorithm. The user will then see the most crucial and beneficial result first [6] and [7].

The search engines will become very successful and popular if they use efficient ranking mechanisms. These days, it is very successful because of its page rank algorithm. Page ranking algorithms are used by the search engines to present the search results by considering relevance, importance, and content score and using web mining techniques to order them according to the user's interest. Some ranking algorithms depend only on the link structure of the documents, i.e., their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining), while some use a combination of both, i.e., they use the content of the document as well as the link structure to assign a rank value for a given document. If the search results are not displayed according to the user's interests, then the search engine will lose popularity. So the ranking algorithms become very important [8] [9].

The PageRank algorithm is based on the hyperlink structure. The Google search engine employs the algorithm PageRank. The most popular algorithm for ranking billions of webpages is the PageRank algorithm. In a query to determine a general ranking score for every webpage, Google's search algorithm mixes pre-calculated PageRank scores with the scores of text matching. The link structure of web pages affects how well the PageRank algorithm works. The PageRank algorithm has been built on the idea that in cases where a page has important links pointing at it, links from that page to another one should also be believed to be important pages. PageRank takes the backlink into account while calculating the rank score. As a result, a page will rank highly if the sum of the ranks of its backlinks is high [10].

The Weighted PageRank algorithm (WPR) assigns rank scores depending on the popularity of pages, taking into account the significance of out-links as well as in-links to pages. WPR outperforms the traditional PageRank algorithm with regard to providing more relevant pages for a given query. The author claims that the more significant a webpage is, the more links to or from other webpages there are likely to be. Rather than splitting a page's rank value equally among its out-link pages, the suggested expanded PageRank algorithm—a Weighted PageRank Algorithm—assigns bigger rank values to more significant (i.e., popular) pages. Each page with an outlink is assigned a value based on its popularity (the number of out- and in-links it has) [11] [12].

According to the Hyperlinked Induced Topic Search algorithm (HITS), each query topic has a set of authoritative sites or pages that are relevant and popularly focused on the topic, as well as hub pages or sites containing helpful links to relevant sites, which include links to several related authorities [13].

This study is divided into six sections. The concept of web mining and its varieties are covered in the first section. The second deals with problem identification, the third with the implementation of the suggested system, the fourth with the experiential outcomes of the suggested system, the fifth with a comparison analysis, and the final section deals with the study's conclusions.

2. Web Mining

Web mining can be defined as the practice of using data mining (DM) approaches in order to search the WWW for knowledge that has been hidden. This knowledge may be found in web page content, links within the WWW, or web server logs. The WWW represents a huge repository of hyperlinked, heterogeneous information that includes images, text, video, audio, and metadata. It is getting harder to manage information on the Web and satisfy user needs as a result of the quick expansion of the sources of information that are available on the Web and the growing needs of users. In reality, we are starved for knowledge while drowning in data. As a result, users are increasingly required to employ information retrieval methods in order to locate, filter, extract, and organize the desired information. Web mining may be divided into three classes, which are: web content mining, web structure mining, and web usage mining. Those categories are determined by the use of the web data utilized as input in the data mining process. The process of knowledge discovery and potentially helpful information from the web are the three categories' main concerns. Although there are three types of web mining, their differences are becoming less distinct because of how closely they are all connected [14].

2.1. *Web Content Mining (WCM)*

WCM can be defined as the process of taking valuable information out of web documents. Images, text, video, audio, and structured data such as lists and tables can all be found in web documents. Since numerous DM methods could be used in WCM, WCM and DM are connected. Since a large portion of the content on the web is text, it is also associated with text mining. Web documents and search engine results pages are both eligible for mining. Either the agent-based method or the database method might be distinguished. The first method tries to enhance information discovery and filtering. The second method seeks to model web data into a more structured form so that it may be analyzed using common database querying techniques and DM applications [15].

2.2. *Web Structure Mining (WSM)*

Web pages and web sites are structurally summarized using WSM. A typical web graph has web pages that act as nodes, and links form edges between pages that are related to one another. WSM seeks to identify the link structures regarding the hyperlinks at inter-document levels, whereas WCM primarily concentrates on the structure of the inner document. The goal of WSM is to identify the model that underlies the Web's link architectures. Whether the link description is present or not, the model depends on the topology of the hyperlink. Also, this model is useful for categorizing web pages and generating data about similarities and connections between web pages. Additionally, the web's link structure carries significant inferred data that could be used to rank or filter web pages. In particular, it is possible to interpret a link from pages A to B as the author of page *A* endorsing page B. It has been suggested that certain new algorithms take advantage of this link structure not just for key-word searches but for other purposes as well, like the automatic creation of Yahoo-like hierarchies or locating online groups. Since such algorithms use more information than just the content of the pages, they typically perform more qualitatively than information retrieval algorithms. Although it is feasible to have a local level in the web's link structure, doing it globally is rather challenging. Therefore, global link analysis algorithms have rather effective anti-spam protections [16].

2.3. *Web Usage Mining (WUM)*

WUM looks for patterns in user navigation in web data and helpful information in secondary data collected from user interactions while browsing the web. It focuses on methods for predicting how users will behave while browsing the web. This type of web mining provides data on web access for webpages. The routes leading to viewed web pages are provided by this usage data. Frequently, the web server's access logs automatically compile this information. Other helpful data is provided by CGI scripts, like user subscription data, referrer logs, and survey logs. The general application of DM by businesses, as well as their intranet- and internet-based applications and information access, depend on this category [17].

3. Problem of the Ranking Algorithms

The following is a summary of the major issues with the discussed ranking algorithms [18]:

- *Rank quality of PageRank:* The discussed algorithms of ranking have demonstrated quite a good quality, and Google's success serves as evidence of this (or the fact that they are still being utilized). But there are certain areas that could be improved.
- *Data Mining Technique of PageRank:* The Web Usage Mining approach, which may greatly enhance the quality of web page rankings according to user information demands, is not used by the PageRank algorithm; instead, it just utilizes Web Content Mining and Web Structure Mining methods.
- *PageRank is static in nature:* The significance, or rank score, that is assigned to every page by the PageRank algorithm is static. Only the web's link structure affects the rank.

4. Proposed System

- In general, the proposed system has two main stages, as illustrated in Figure 1:

Stage 1: Link’s Metadata Scraping.
Stage 2: Semantic Metadata Analysis.

4.1. *Link’s Metadata Scraping Stage*

A Google-programmable customized search engine (CSE) will be designed and implemented to extract and store the metadata of the links in JSON format. The related links

will be automatically ranked using Google's PageRank algorithm. In general, the first stage has four main steps, as listed below:
Step 1: Create a programmable CSE based on an XML file.
Step 2: Design and implement the search box in the XML file.
Step 3: Determine the API key for the programmable CSE.
Step 4: Extract metadata from the programmable CSE using JSON API.

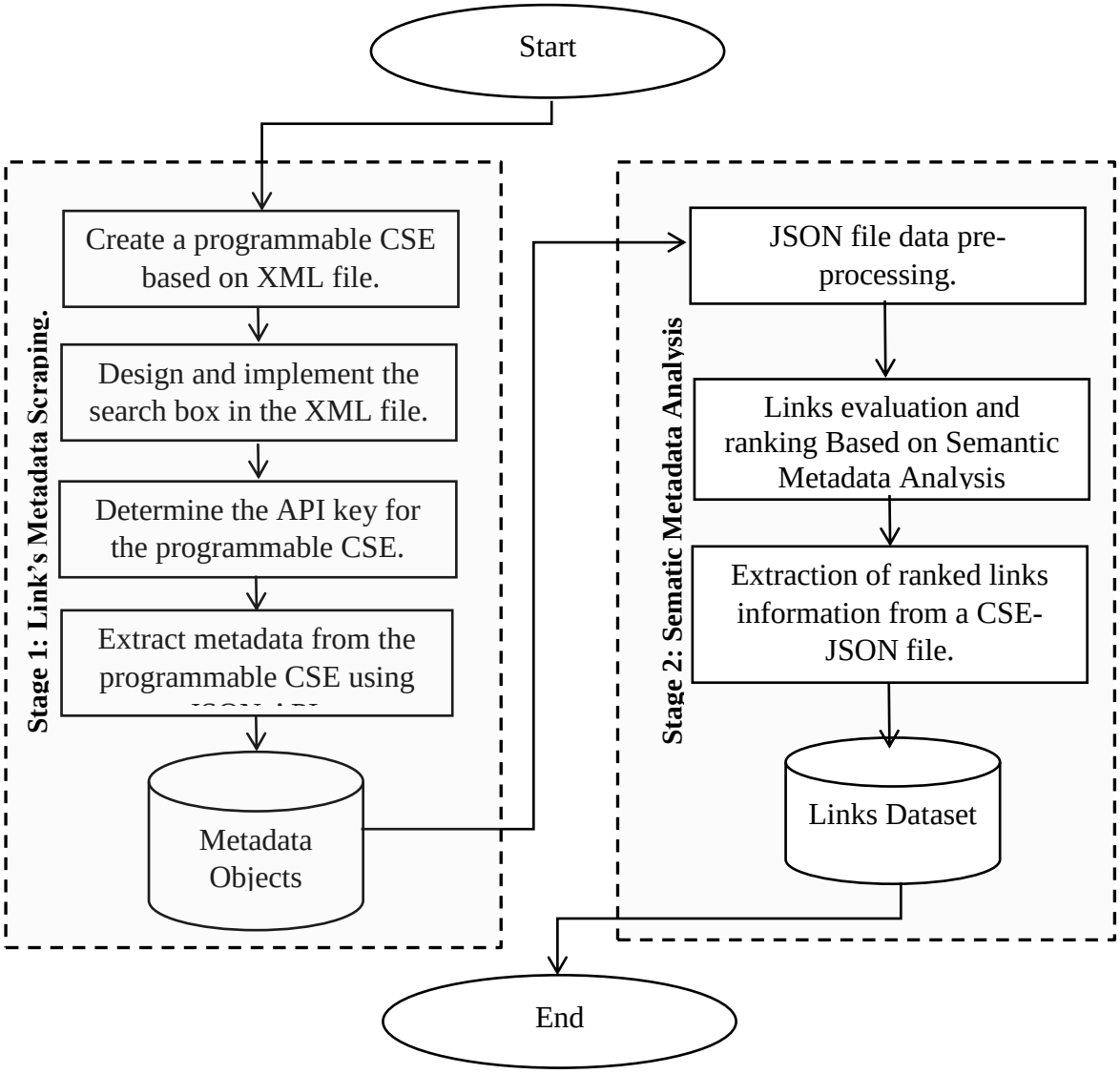

Figure 1: Block diagram of the proposed system

4.1.1. *Create a Programmable CSE Based on XML File*

In this step, the programmable CSE based on a JSON file using an API is defined and created. Once the API key is defined and created, the programmable CSE and the XML file for the programmable CSE are downloadable, as are the annotations and context of the CSE-XML file. In general, a programmable CSE based on an XML file has two components. Each is controlled by the XML file designed and constructed to run and control the programmable CSE. The first component (context) of the XML file is the control tag. This context in the XML file describes the basic features of the programmable CSE, such as wherever using the image searching option or promotions in the programmable CSE is enabled. The XML file's context outlines the search engine's structure and establishes its behavior. The second

component of the XML file that describes and runs the programmable CSE is the annotations tag, which includes which webpages or websites (indicating preferences) that the CSE wants to be included in the search engine. The programmable CSE created using an XML file is illustrated in algorithm (1).

Algorithm (1): XML file configuration

Input: Initial XML file**.**
Output: XML file format**.**
Begin
Step 1: Create the root element.
Step 2: Create annotation tag.
Step 3: Associate the website with its search engine according to the following steps:
Step 3.1: Configure the search label tag.
Step 3.2: Specify the way that the site must be treated by search engine through getting labels for search engine from Advanced tab's Context section in Control Panel.
Step 3.3: Set the search engine label's name within context file. The label for including the sites is in a form of _cse_xxxxxxxxxxx, where x represents a character, and the label for the exclusion of the sites is in a form of the _cse_exclude_xxxxxxxxxxx.
Step 4: for the addition of more sites, another annotation element should be created and defined.
Step 5: Save and export XML file.
End

4.1.2. *Design and Implement the Search Box in the XML File*

The second step of the first stage is implementing the search engine query after the programmable CSE is created. In this step, the element of the programmable CSE is added to the XML file. To do that, some code must be copied and pasted into the HTML file where the search engine is called. The following steps illustrate the main algorithm (2) used to implement the search query in the programmable CSE.

Algorithm (2): CSE Query

Input: XML file, CSE- ID, user query**.**
Output: HTML page (XML code).
Begin
Step 1: In the Control Panel, click on the search engine that is used as a programmable CSE.
Step 2: On the sidebar, click the Setup section, then the Basic tab.
Step 3: In the details section, click on the Get code section and paste the designed code into the HTML page source code where the programmable CSE Element appears.
End

4.1.3. *Determine the API Key for the Programmable CSE*

The programmable CSE requires an API key. Google can identify the client by using the API key. The user can ask the current search engine to do requests with it, such as importing the link data, performing searches, exporting the link data, and other operations, by requesting access to the API. The following algorithm (3) must be used in order to determine the API ID number before using the API in the programmable CSE.

Algorithm (3): API Key Request

Input: CSE name, CSE-ID**.**
Output: API key.
Begin
Step 1: Create the programmable CSE-based API key.
Step 1.1: Use the Control Panel and sign in using your Google account.
Step 1.2: In the Sites to Search section, include the programmable CSE that needs to access the database. The URL pattern can be included in the XML file.
Step 1.3: From the programmable CSE Navigate, click on Add to request a new API key.
Step 1.4: Select the configuration file of the programmable CSE Navigate by clicking on Adding to request a programmable CSE and clicking on the Create button.
Step 1.5: confirm the configuration file of the programmable CSE and extend it to search the entire web.
S Step 2: Enable the entire webpage search by modifying the programmable CSE.
Step 2.1: Delete the existing URL, and toggle on the Search the entire web option.
Step 3: Get your search engine API key.
Step 3.1: An API key is needed to use the Tailored Searching JSON API.
Step 3.2: Navigate to the customized searching JSON API page and click Get a Key.
Step 3.3: The search engine ID can be found at the end of the page.
Step 3.4: Choose an existing project or create a new one and click Next.
Step 4.4: Copy and export your API key.
End

4.1.4. *Extract Metadata from the Programmable CSE Using JSON API*

The JSON file will be extracted in this step. An application may be developed to retrieve and display search results from the programmable CSE using the JSON file that provides the link's metadata. The API allows us to submit a full request to obtain a JSON file format containing the link's metadata. A JSON file is imported as a result of a search query after the search query has been implemented on a programmable CSE. The JSON file typically contains three categories of metadata:

- Metadata that describes the requested search and, potentially associated search requests.
- Metadata that describes the search results.
- Metadata that describes a programmable search engine.

4.2. *Semantic Metadata Analysis Stage*

In this stage, after a JSON file that contains the metadata of the links has been extracted from the first stage, a new ranking method can be used for the evaluation of the importance of links and re-rank them based upon semantic metadata criteria across almost all time periods, regions, and related topics and queries. In general, the second stage has three main steps, as listed below:

Step 1: JSON file data pre-processing.
Step 2: Links evaluation and ranking.
Step 3: Extraction of ranked links information from a CSE- JSON file.

4.2.1. *JSON File Data Preprocessing*

In this step, the JSON file data that was collected in the first stage is processed. Once this step is implemented, the JSON file is opened in write mode, and the json.dump () function is used for serializing the Python dictionary as a JSON-formatted stream to the opened file. In

the case of data that contains non-ASCII characters in the JSON file, JSON data file cleaning is designed and implemented. In order to achieve that, ensure-ASCII = False must be passed to the json.dump() function. After that, the json.load() function will automatically return a Python dictionary, which eases our work with JSON files. The following steps illustrate the main algorithm (4) used to implement JSON file data preprocessing.

Algorithm (4): JSON File Data Preprocessing

Input: JSON file.
Outputs: Processed JSON file.
Begin
Step 1: Import the JSON file library in Python using import JSON.
Step 2: Design the Python dictionary to be saved as a JSON file.
Step 3: Utilize open () function for opening the JSON file in write mode.
Step 4: Use json.dump () function to serialize Python dictionary as a JSON formatted stream to the opened file.
Step 5: Clean JSON file data and handle non-ASCII characters.
Step 6: Use json.load () function to automatically return a Python dictionary and eases our work with JSON files.
End.

4.2.2. *Links Evaluation and Ranking*

After a programmable CSE has been designed and implemented to extract the metadata of the links and automatically rank them using Google's PageRank algorithm, a semantic metadata analysis will be used to evaluate the importance of the links and re-rank them related to the given query in order to extract the most significant top 10 links that have been ranked based on five criteria:

- The number of links visited yearly.
- The number of links visited hourly.
- The number of links visited by region.
- The number of links visited by related topics.
- The number of links visited by related queries.

A client-side script will be utilized for counting hits or visits on the links. Whenever a link has been accessed, a script will be loaded on the client side from the web server. The script will be monitoring clicks in addition to any keyboard events that occur. In the case where an event occurs and that event happens over a link, then it will send a message to the web server with information about the current link. A database of log files will be used on the server side to record link IDs and hit counts. The hit count will be incremented whenever a hit occurs on the link. The database or log files will be accessed by the crawler at crawl time. This crawled information (i.e., the hit count) will be stored in the search engine's database, which is utilized for the calculation of rank values for various links. The following steps illustrate the main algorithm (5) used to implement link ranking using semantic metadata analysis.

Algorithm (5): Links Evaluation and Ranking

Input: Processed JSON file.
Outputs: A list of the top 10 ranked links from each of semantic analysis criteria.
Begin
Step 1: Create a trend request object to access each link's metadata by using the pytrends.
Step 2: Evaluate and rank the links in the processed JSON file using semantic analysis criteria.
Step 2.1: Use the get the interest over time method to evaluate the importance of the links and re-rank them based upon the number of links visited yearly.
Step 2.2: Use the get hourly historical interest method to evaluate importance of links and re-rank them based upon the number of links visited hourly.
Step 2.3: Use the get the interest by region method to evaluate importance of links and re-rank them based upon the number of links visited by region.
Step 2.4: Use the get related topics method to evaluate importance of links and re-rank them based upon the number of links visited by related topics.
Step 2.5: Use the get related queries method to evaluate importance of links and re-rank them based upon number of links visited by related queries.
Step 3: Get the top 10 ranked link from each of the five semantic analysis criteria.
Step 4: Extract the top-ranked link from each of the five semantic analysis criteria.
End

4.2.3. *Extraction of Ranked Links Information from a CSE-JSON File*

In this step, after the link's metadata dataset has been collected using our designed programmable CSE and the links have been evaluated and re-ranked using semantic metadata criteria across almost all time periods, regions, and related topics and queries, the information about the ranked links will be extracted from the CSE-JSON file. The following steps illustrate the main algorithm (6) used to extract the information about the ranked links from a CSE-JSON file.

Algorithm (6): Extraction of Ranked Links Information

Input: A list of the top 10 ranked links.
Output: The information about the ranked links
Begin
Step 1: Get the result items from the top 10 ranked links list.
Step 2: For each of the top 10 ranked links.
Step 2.1: Extract the title of the link from the CSE-JSON file.
Step 2.2: Extract the snippet of the link from the CSE-JSON file.
Step 2.3: Extract the HTML snippet of the link from the CSE-JSON file.
Step 2.4: Extract the URL of the link from the CSE-JSON file.
Step 2.5: End for.
Step 3: display the links and the information about them at top of the result list.
End

5. The Experiential Results of the proposed System

The proposed system consist of two stages:
Stage 1: Link's Metadata Scraping.
Stage 2: Links Evaluation and Ranking.

5.1. *Link's Metadata Scraping Stage*

After the search query is implemented by the user, such as for information retrieval, a JSON file is imported as a result of the search query. The JSON file that contains the metadata of the links lets the user retrieve and display search results, as shown in Table 1.

Table 1: The Metadata of the Links

Rank	URL	Metadata
1	https://www.google.com/patents/US9405794	Title: US9405794B2 – Information retrieval system. Description: A system of information retrieval is responsible for the conversion of the unstructured ad-hoc search queries to structured search instructions retrieving the data in structured…
2	https://www.google.com/patents/US7783643	Title: US7783643B2 – Direct navigation for information retrieval. Description: an approach of the document retrieval has been provided in this study, this approach includes the assignment of the concept labels to the documents that are contained in the collection based on the rules of grammar…
3	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While the Research Information keeps maturing as an expertise area, discussions about implementing and adopting of standardization initiatives, like CERIF and CASRAI, had intensified…
4	https://www.google.com/patents/US5784608	Title: US5784608A – Hypertext information retrieval using profiles and… Description: A computer-implemented approach and system for the retrieval of the information. A first information file is received, including a first markup language to…
5	https://google.com/patents/US20140105467	Title: US20140105467A1 – Image Classification And Information Retrieval… Description: … of digital facial images received over wireless digital networks or Internet and information retrieval that is related to the classified image.
6	https://www.google.com/patents/US20150293900	Title: US20150293900A1 – Information retrieval system based on a… Description: Embodiment of invention provides the methods and the systems for the representation of plurality of the languages in lexicon based upon a unified language model.
7	https://www.google.com/patents/US6385605	Title: US6385605B1 – Information retrieval apparatus and a method… Description: An apparatus of information retrieval performs the retrieval of the information from a data-base. In the case where a request for the retrieval has been received from the user, a first section of retrieval…
8	https://www.google.com/patents/US6611834	Title: US6611834B1 – Customization of information retrieval through user… Description: the user at the client machine has the ability of customizing the components of a data-base search that is performed at the server. The user can do this through sending the executable code to…
9	https://www.google.com/patents/US3744030	Title: US3744030A – Intrinsic controls for information retrieval systems… Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, utilizing digital code that has been carried out of series of the chosen…
10	https://www.google.com/patents/US7702618	Title: US7702618B1 – Information retrieval system for archiving multiple… Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of the documents. Phrases have been identified which predict presence of other…

5.2. *Links Evaluation and Ranking Stage*

In this stage, after a JSON file that contains the metadata of the links has been extracted and these links have been automatically ranked using Google's PageRank algorithm, the importance of the links will be evaluated and re-ranked related to the given query in order to extract the most significant top 10 links from each of the five semantic analysis criteria:

Criteria 1: Links can be ranked based on the number of links visited annually, as shown in Table 2.

Table 2: Ranked Links Based on Interest over Time

Rank	URL	Metadata	Hits
1	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval approach has been provided in this study. It includes the assignment of concept labels to the documents that are contained in the collection based on the rules of the grammar...	100
2	https://www.google.com/patents/US9405794	Title: US9405794B2 - Information retrieval system. Description: A system of information retrieval performs the conversion of the unstructured ad-hoc search queries to structured instructions of the search retrieving the data in structured...	74
3	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: invention embodiments provide the methods and systems for the representation of plurality of the languages in lexicon based upon one unified model of the language.	40
4	https://google.com/patents/US20140105467	Title: US20140105467A1 - Image Classification And Information Retrieval... Description: ... of digital facial images that are received over the wireless digital networks or via Internet and retrieval of the information related to the classified images.	26
5	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs digital code that is formed of series of the chosen...	20
6	https://www.google.com/patents/US7702618	Title: US7702618B1 - Information retrieval system for archiving multiple... Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of the documents. Phrases have been identified which predict existence of other...	18
7	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While the Research Information keeps maturing as an expertise area, discussions about adoption and implementation of the initiatives of the standardization, like CERIF and CASRAI, had intensified...	14
8	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first	12

		file of the information has been received, including a first markup language to...	
9	https://www.google.com/patents/US6385605	Title: US6385605B1 - Information retrieval apparatus and a method... Description: An apparatus of information retrieval performs the retrieval of the information from data-base. In the case where a request of the retrieval has been received from the user, first retrieval section...	11
10	https://www.google.com/patents/US6611834	Title: US6611834B1 - Customization of information retrieval through user... Description: A user at client machine has the ability of customizing the components of data-base search that is carried out at the server. The user performs that through sending the executable code to...	10

Criteria 2: Links can be ranked based on the number of links visited hourly, as shown in Table 3.

Table 3: Ranked Links Based on Hourly Historical Interest

Rank	URL	Metadata	Hits
1	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs digital code that has been formed of a set of the chosen...	73
2	https://www.google.com/patents/US7702618	Title: US7702618B1 - Information retrieval system for archiving multiple... Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of the documents. Phrases have been identified predicting existence of other...	70
3	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While Research Information keeps maturing as an expertise area, discussions about implementations and adoptions of the initiatives of standardization, like CERIF and CASRAI, had intensified...	69
4	https://www.google.com/patents/US6611834	Title: US6611834B1 - Customization of information retrieval through user... Description: A user at client machine has the ability of customizing the components of data-base search that has been carried out at the server. The user performs that through the sending of the executable code to...	68
5	https://www.google.com/patents/US6385605	Title: US6385605B1 - Information retrieval apparatus and a method... Description: An apparatus of the information retrieval retrieves the information from data-base. In the case where a request for retrieval has been received from the user, first retrieval section...	63
6	https://www.google.com/patents/US9405794	Title: US9405794B2 - Information retrieval system. Description: A system of information retrieval performs the conversion of the unstructured ad-hoc search queries to structured search instructions retrieving the data in structured...	52
7	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval approach has been provided in this study. This approach includes the	44

		assignment of the concept labels to the documents that are contained in collection based on the grammar rules...	
8	https://google.com/patents/US20140105467	Title: US20140105467A1 - Image Classification And Information Retrieval... Description: ... of digital facial images received via wireless digital networks or Internet and retrieving information that is related to the classified image.	42
9	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: Embodiments of invention provide the methods and the systems for the representation of several languages in lexicon based upon one unified language model.	34
10	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first information file is received, including first markup language to...	28

Criteria 3: Links can be ranked based on the number of links visited by region, as shown in Table 4.

Table 4. Ranked Links Based on Interest by Region

Rank	URL	Metadata	Hits
1	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first information file has been received, including first markup language to...	100
2	https://www.google.com/patents/US9405794	Title: US9405794B2 - Information retrieval system. Description: A system of information retrieval performs the conversion of the unstructured ad-hoc search queries to structured search instructions retrieving the data in structured...	79
3	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: Invention embodiments provide the systems and approaches for the representation of several languages in a lexicon based upon unified language model.	36
4	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs a digital code that has been formed of series of the chosen...	31
5	https://www.google.com/patents/US7702618	Title: US7702618B1 - Information retrieval system for archiving multiple... Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of documents. Phrases have been identified predicting existence of other...	30
6	https://www.google.com/patents/US6385605	Title: US6385605B1 - Information retrieval apparatus and a method... Description: An apparatus of information retrieval performs the retrieval of the information from data-base. In the case where a request of retrieval has been	28

		received from a user, first section of retrieval ...	
7	https://google.com/patents/US20140105467	Title: US20140105467A1 - Image Classification And Information Retrieval... Description: ... of the digital facial images that have been received via the wireless digital networks or Internet and retrieval of information that is related to image.	25
8	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval method has been provided. This approach includes the assignment of concept labels to documents that are contained in a collection based on the rules of grammar...	22
9	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While Research Information keeps maturing as an expertise area, discussions about implementations and adoptions of initiatives of standardization, like CERIF and CASRAI, had intensified...	20
10	https://www.google.com/patents/US6611834	Title: US6611834B1 - Customization of information retrieval through user... Description: The user at the client machine has the ability of customizing the components of data-base search that is performed at the server. The user can do that through the sending of the executable code to ...	17

Criteria 4: Links can be ranked based on the number of links visited by related topics of a keyword, as shown in Table 5.

Table 5: Ranked Links Based on Related Topics

Rank	URL	Metadata	Hits
1	https://www.google.com/patents/US6385605	Title: US6385605B1 - Information retrieval apparatus and a method... Description: An apparatus of information retrieval performs the retrieval of the information from data-base. In the case where a request of retrieval has been received from a user, first section of retrieval ...	100
2	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs a digital code that has been formed of series of the chosen...	99
3	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: Invention embodiments provide the systems and approaches for the representation of several languages in a lexicon based upon unified language model.	66
4	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first information file has been received, including first markup language to...	30
5	https://www.google.com/patents/US7702618	Title: US7702618B1 - Information retrieval system for archiving multiple... Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of documents. Phrases	17

		have been identified predicting existence of other...	
6	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval method has been provided. This approach includes the assignment of concept labels to documents that are contained in a collection based on the rules of grammar...	10
7	https://google.com/patents/US20140105467	Title: US20140105467A1 - Image Classification And Information Retrieval... Description: ... of the digital facial images that have been received via the wireless digital networks or Internet and retrieval of information that is related to image.	5
8	https://www.google.com/patents/US9405794	Title: US9405794B2 - Information retrieval system. Description: A system of information retrieval performs the conversion of the unstructured ad-hoc search queries to structured search instructions retrieving the data in structured...	4
9	https://www.google.com/patents/US6611834	Title: US6611834B1 - Customization of information retrieval through user... Description: The user at the client machine has the ability of customizing the components of data-base search that is performed at the server. The user can do that through the sending of the executable code to ...	3
10	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While Research Information keeps maturing as an expertise area, discussions about implementations and adoptions of initiatives of standardization, like CERIF and CASRAI, had intensified...	2

Criteria 5: Links can be ranked based on the number of links visited by related search queries, as shown in Table 6.

Table 6: Ranked Links Based on Related Search Queries

Rank	URL	Metadata	Hits
1	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: Invention embodiments provide the systems and approaches for the representation of several languages in a lexicon based upon unified language model.	100
2	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs a digital code that has been formed of series of the chosen...	71
3	https://www.google.com/patents/US9405794	Title: US9405794B2 - Information retrieval system. Description: A system of information retrieval performs the conversion of the unstructured ad-hoc search queries to structured search instructions retrieving the data in structured...	70
4	https://google.com/patents/US20140105467	Title: US20140105467A1 - Image Classification And Information Retrieval... Description: ... of the digital facial images that have been received via the wireless digital networks or Internet and retrieval of information that is related to image.	69
5	https://www.google.com/pate	Title: US6385605B1 - Information retrieval apparatus and	68

	nts/US6385605	a method... Description: An apparatus of information retrieval performs the retrieval of the information from data-base. In the case where a request of retrieval has been received from a user, first section of retrieval ...	
6	https://www.google.com/patents/US7702618	Title: US7702618B1 - Information retrieval system for archiving multiple... Description: A system of information retrieval utilizes the phrases for the indexing, retrieval, organizing and description of documents. Phrases have been identified predicting existence of other...	57
7	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval method has been provided. This approach includes the assignment of concept labels to documents that are contained in a collection based on the rules of grammar...	48
8	https://www.google.com/mymaps/viewer?mid=1zgqgaOn9U8Uqe1TvE6huiQP8MoE&hl=en_US	Title: Research papers on information retrieval systems. Description: While Research Information keeps maturing as an expertise area, discussions about implementations and adoptions of initiatives of standardization, like CERIF and CASRAI, had intensified...	44
9	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first information file has been received, including first markup language to...	37
10	https://www.google.com/patents/US6611834	Title: US6611834B1 - Customization of information retrieval through user... Description: The user at the client machine has the ability of customizing the components of data-base search that is performed at the server. The user can do that through the sending of the executable code to ...	28

After the links have been evaluated and re-ranked using five semantic metadata criteria, the top-ranked link from each of the five semantic analysis criteria will be extracted in order to display the most valuable pages according to the user's query at the top of the result list, as shown in Table 7.

Table 7: The Top-Ranked Link from Each of the Five Semantic Analysis Criteria.

No.	URL	Metadata
1	https://www.google.com/patents/US7783643	Title: US7783643B2 - Direct navigation for information retrieval. Description: A document retrieval method has been provided. This approach includes the assignment of concept labels to documents that are contained in a collection based on the rules of grammar...
2	https://www.google.com/patents/US3744030	Title: US3744030A - Intrinsic controls for information retrieval systems... Description: Intrinsic controls have been provided for specific automatic functions in a system of information retrieval, which employs a digital code that has been formed of series of the chosen...
3	https://www.google.com/patents/US5784608	Title: US5784608A - Hypertext information retrieval using profiles and... Description: A computer-implemented approach and system for the retrieval of the information. A first information file has been received, including first markup language to...
4	https://www.google.com/patents/US6385605	Title: US6385605B1 - Information retrieval apparatus and a method... Description: An apparatus of information retrieval performs the

		retrieval of the information from data-base. In the case where a request of retrieval has been received from a user, first section of retrieval ...
5	https://www.google.com/patents/US20150293900	Title: US20150293900A1 - Information retrieval system based on a... Description: Invention embodiments provide the systems and approaches for the representation of several languages in a lexicon based upon unified language model.

6. Comparison of Proposed System with Related Works

From Table 8, it is clear that the proposed system is much better than other ranking algorithms.

Table 8: Comparison of the Proposed System with Related Works.

Parameter	PageRank (PR)	Weighted Page Rank (WPR)	Hyperlinked Induced Topic Search (HITS)	Proposed System
Description	Computes scores at the indexing time. Results are arranged in order of page importance.	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes hub and authority scores of the relevant pages on the fly. Relevant as well as important pages are returned.	Computes scores at indexing time. Pages are sorted according to importance and relevance.
Mining Technique Used	Web Structure Mining.	Web Structure Mining.	Web content Mining, Web Structure Mining	Web Usage Mining, Web Structure Mining
Complexity	O(log n)	< O(log n)	< O(log n)	> O(log n)
I/P Parameters	Inbound links of pages.	Inbound links and Outbound links of pages.	Content, Inbound links and Outbound links of pages.	Visit Counts of links.
Relevancy of pages	No	No	Yes	Yes
Importance of pages	Yes	Yes	No	Yes
Quality of result	Low	High	low	High
Advantages	Computation of ranks using the least amount of complexity and effort.	Computation of ranks with minimum effort and less complexity.	It is sensitive to the user query. Computes the authority and hubs correctly.	As user feedback is taken into consideration, the pages that are returned are of a high relevancy and quality. As pages are ordered based on users' information demands, search space could be greatly reduced.
Limitations	In calculating rank, pages' relevance is not taken into account. Links are all given the same level of importance.	No relevancy of pages is considered in rank computation. All links are considered equally important.	Topic drift and Efficiency problems occur. Non-relevant documents can be retrieved.	Extra work needed on the part of crawlers to retrieve page visit counts from web servers.

7. Conclusions

The challenge of extracting useful information from a vast amount of data is quite difficult, and the World Wide Web is essential for gathering and sharing information. The search for relevant information is done extremely effectively using the ranking algorithms. Various approaches employ different ranking algorithms. Using a new ranking method that can be used to evaluate the importance of the links based on semantic metadata analysis, this method

provides more relevant ranking results than other ranking methods. Therefore, once a link is ranked highly by this ranking method, it is guaranteed to contain important information related to the given query, as this method ranks links based on the meaning of the links to stratify their importance, not the number of in-links the pages have, avoiding meaningless links from being scored highly, which is a problem with other ranking methods. The suggested system consists of two main stages. In the first stage, the metadata of the links can be extracted and automatically ranked using Google's programmable CSE. In the second stage, the importance of the links can be evaluated and re-ranked based on the number of links visited across almost all time periods, regions, and related topics and queries. The suggested system uses the user's query to find more relevant information. Thus, this idea is particularly helpful for placing the most valuable pages at the top of the result list based on user browsing behavior, thus reducing the search space.

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] M. K. Alshammery and A. F. Aljuboori, "Classifying Illegal Activities on Tor Network using Hybrid Technique," *Iraqi Journal of Science*, vol. 63, no. 9, pp. 3994–4004, 2022, doi: 10.24996/ijs.2022.63.9.30.

[3] M. K. Alshammery and A. F. Aljuboori, "Crawling and Mining the Dark Web: A Survey on Existing and New Approaches," *Iraqi Journal of Science*, vol. 63, no. 3, pp. 1339–13487, 2022, doi: 10.24996/ijs.2022.63.3.36.

[4] M. N. A. Khan and A. Mahmood, "A distinctive approach to obtain higher page rank through search engine optimization," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 43, no. 3, pp. 1–12, 2018, doi: 10.1007/s12046-018-0812-3.

[5] D. Sharma, R. Shukla, A. K. Giri, and S. Kumar, "A brief review on search engine optimization," *Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2019*, pp. 687–692, 2019, doi: 10.1109/CONFLUENCE.2019.8776976.

[6] Markus Fällman, "Analysing a modified ranking algorithm for exploratory search," *Master's thesis in Engineering Mathematics and Computational Science, Chalmers University of Technology Gothenburg*, Sweden, 2020.

[7] R. Gao and C. Shah, "Toward creating a fairer ranking in search engine results," *Inf. Process. Manag.*, vol. 57, no. 1, p. 102138, 2020, doi: 10.1016/j.ipm.2019.102138.

[8] S. Goel, R. Kumar, M. Kumar, and V. Chopra, "An efficient page ranking approach based on vector norms using sNorm(p) algorithm," *Inf. Process. Manag.*, vol. 56, no. 3, pp. 1053–1066, 2019, doi: 10.1016/j.ipm.2019.02.004.

[9] F. Ali and S. Khusro, "Content and link-structure perspective of ranking webpages: A review," *Comput. Sci. Rev.*, vol. 40, p. 100397, 2021, doi: 10.1016/j.cosrev.2021.100397.

[10] R. Mathur, V. Pathak, and D. Bandil, *Improved Google Page Rank Algorithm*, vol. 841. Springer Singapore, 2019. doi: 10.1007/978-981-13-2285-3.

[11] F. Alhaidari, S. Alwarthan, and A. Alamoudi, "User Preference Based Weighted Page Ranking Algorithm," *ICCAIS 2020 - 3rd Int. Conf. Comput. Appl. Inf. Secur.*, pp. 1–6, 2020, doi: 10.1109/ICCAIS48893.2020.9096823.

[12] H. Alghamdi and F. Alhaidari, "Extended User Preference Based Weighted Page Ranking Algorithm," *Proc. - 2021 IEEE 4th Natl. Comput. Coll. Conf. NCCC 2021*, pp. 2–7, 2021, doi: 10.1109/NCCC49330.2021.9428844.

[13] G. A. Al-Sultany and I. K. Abbood, "Conferences events suggestion using ranked based hyperlink-induced topic search algorithm," *J. Comput. Theor. Nanosci.*, vol. 16, no. 3, pp. 982–988, 2019, doi: 10.1166/jctn.2019.7987.

[14] M. J. Hamid 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, doi: 10.14569/IJACSA.2018.090630.

[15] D. Vats and A. Sharma, "Analysis and comparison of various web mining techniques," *J. Comput. Theor. Nanosci.*, vol. 16, no. 10, pp. 4125–4134, 2019, doi: 10.1166/jctn.2019.8491.

[16] A. P. Phyu and E. E. Thu, "Short Survey Of Data Mining And Web Mining Using Cloud Computing," *Int. J. Adv. Netw. Appl.*, vol. 12, no. 05, pp. 4725–4731, 2021, doi: 10.35444/ijana.2021.12509.

[17] K. Griazev and S. Ramanauskaite, "Web mining taxonomy," *2018 Open Conf. Electr. Electron. Inf. Sci. eStream 2018 - Proc.*, pp. 1–4, 2018, doi: 10.1109/eStream.2018.8394124.

[18] A. Monelli and S. B. Sriramoju, "An overview of the challenges and applications towards web mining," *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 127–131, 2019, doi: 10.1109/I-SMAC.2018.8653669.