ISSN: 0067-2904

Link Analysis in Web Information Retrieval: a Survey

Matheel E. Abdulmunem, Esraa Q. Naamha*
Department of Computer Science, Technology University, Baghdad, Iraq

Received: 17/5/2022 Accepted: 23/10/2022 Published: 30/9/2023

Abstract

The analysis of the hyperlink structure of the web has led to significant improvements in web information retrieval. This survey study evaluates and analyzes relevant research publications on link analysis in web information retrieval utilizing diverse methods. These factors include the research year, the aims of the research article, the algorithms utilized to complete their study, and the findings received after using the algorithms. The findings revealed that Page Rank, Weighted Page Rank, and Weighted Page Content Rank are extensively employed by academics to properly analyze hyperlinks in web information retrieval. Finally, this paper analyzes the previous studies.

Keywords: Link Analysis, Web Data, Information Retrieval, Ranking Algorithms.

استرجاع معلومات الويب من خلال تحليل الروابط التشعبية: دراسة استقصائية

مثيل عماد الدين عبد المنعم، اسراء قاسم نعمة *

قسم عموم الحاسوب، الجامعة التكنولوجية، بغداد، العراق.

الخلاصة

ان تحليل بنية الارتباط التشعبي للوب ادى إلى تحسينات كبيرة في استرجاع معلومات الوب . تقوم هذه الدراسة الاستقصائية بتقييم وتحليل العديد من الدراسات والبحوث المتاحة في مجال استرجاع معلومات الوب من خلال تحليل الروابط باستعمال طرق متنوعة. تشمل هذه الدراسة مقارنة تتضمن سنة البحث ، وأهداف البحث ، والخوارزميات المستعملة لإكمال ابحاثهم ، والنتائج التي تم الحصول عليها بعد استعمال الخوارزميات. كشفت النتائج أن Page Rank و Weighted Page Rank و Weighted Page Content Rank تم توظيفها على نطاق واسع من قبل الأكاديميين لتحليل الروابط التشعبية بشكل صحيح في استرجاع معلومات الوب. أخيرًا ، قام الباحثون بتحليل وعمل مقارنة بين الدراسات السابقة.

1. Introduction

Information retrieval (IR) has made great progress in recent years, and industry professionals and research groups have predicted a bright future for the area [1]. The generic Knowledge Retrieval job specifies how much knowledge a user needs in a given context to solve a problem. The IR model [2] consists of a collection of assumptions and an algorithm for evaluating articles in response to a user query. The discipline of information retrieval has risen in importance in recent years as a result of the intriguing difficulties raised by using the Internet and the World

*Email: esraa.noony@gmail.com

Wide Web as an infinite repository of knowledge. The popularity of web search engines [1][2] backs this up.

Users are becoming increasingly reliant on search engines' ranking techniques to locate material relevant to their needs due to the huge expansion of the Web and the greater expectation put on search engines to anticipate and infer their information demands [1][2].

Users anticipate discovering information in the top-ranked results, and they frequently skim the first few result pages for document samples before exiting or rephrasing their question. This may result in considerable bias in their information search, necessitating ranking algorithms that consider not only the overall page quality and relevancy to the query but also the match with the users' true search intent while constructing the inquiry. Context-aware search services, as well as introducing context into current search services, have the potential to increase retrieval effectiveness and eliminate biases in Web information access [1][2].

The World-Wide-Web (often referred to as the Web) is a massive information database in which Uniform Resource Locator (URL) and hypertext links are used to identify documents and other resources (such as audio, video, photos, and metadata)[3]. Because the Internet is always expanding and changing, it is challenging to keep up with the quantity of information available. As a result, efficient information retrieval systems to locate and arrange required data have become critical. Link analysis has been successfully utilized to determine which web pages to add to the document collection (i.e., which pages to crawl) and organize documents that match a user query (i.e., how to rank pages).

It's also been used to classify online pages, find duplicate sites, and solve several other web information retrieval issues [2][3]. Furthermore, link analysis is the process of building connections between different entities, such as customers and other customers or customers and items. Relationship analysis illustrates how other variables or attributes might be used to characterize the nature and strength of the link. For mapping and analyzing spheres of influence, link analysis is critical [4].

The purpose of this survey is to examine and assess previous research publications on link analysis in web information retrieval using diverse methodologies from four different perspectives:

- The research year.
- The objectives behind the research paper.
- The algorithms used to accomplish their research.
- The results that were obtained after applying the algorithms.

The remainder of this survey paper is organized as follows: Section 2 presents many previous papers related to link analysis in web information retrieval using different algorithms; Section 3 analyses studies; and Section 4 concludes the survey paper.

2. Related Work

Spoerri et al. [5] present the Info Crystal that was developed to assist consumers in searching for information more effectively. First, consider how to depict all of the various relationships between concepts. Propose the design of an Info Crystal with selectable elements that may be displayed to highlight the qualitative or quantitative data associated with them. They've also developed a layout algorithm that allows them to build Info Crystals from inputs; the purpose of this algorithm is to generate an internal icon layout that ensures none of the icons are in the

same position. Call it the rank layout principle, since it precisely follows the rank coding concept.

De Bra et al. [6] present hypertext systems that provide a limited view of the jumps that a reader could want. The only jumps allowed are those that follow links, which are hard-wired. Full-text search and/or index building are only possible when the hypertext system has complete control over the hypertext. They employ a strategy that reduces a hypertext's structure to a hierarchy and a collection of cross-reference links. The key issues in executing information retrieval on a distributed hypertext are: identifying a decent starting point for the search; determining the "optimal" order in which to retrieve nodes. They proposed a navigational algorithm called the "Fish-Search." Only the second point is addressed by the algorithm. The presented algorithm does a partial search over a distributed hypertext using heuristics that have been validated by a large number of simulations. The algorithm was built on top of "Mosaic for X", a popular "World-Wide Web" browser. It converts the browser into a World-Wide-Web information retrieval tool by searching the contents (body) of documents on the Web.

Henzinger et al. [2] propose two successful link analysis techniques to analyze hyperlinks in web information retrieval. The two strategies are Query-Independent Connectivity-Based Ranking and Query-Dependent Connectivity-Based Ranking. Without a specific user query, each page is given a score in a query-independent ranking, with the goal of evaluating the intrinsic quality of the page. With or without specific query-dependent parameters, this score is utilized to rank all articles matching the query at query time. Some of the sites in the query-dependent ranking are given a score that measures the quality and relevance of a page to a specific user query. The findings suggest that these methods are beneficial for assessing hyperlinks in web information retrieval.

Ding et al. [7] stated that the World Wide Web is the proliferation of hyperlinks between web pages that allow a person to navigate from one web page to another with a single click. The Web may be represented as a directed graph, with web pages acting as nodes and hyperlinks acting as directed edges. This graph of hyperlinks contains the following information: If web page pi contains a link to web page pj, it usually means that the creator of pi believes pj contains information that is useful to pi. As a result, such useful ideas and knowledge are recorded as hyperlinks. The Hyperlink Induced Topic Search (HITS) method developed by Kleinberg expands on the basic principles of hubs and authority. HITS assigns significance rankings to hubs and authorities.

Amitay et al. [8] present numerous elements and uses of temporal data in the context of Web Information Retrieval (Web IR). They imagine and present an ideal scenario in which search engines track and record temporal data for each page in their database. Assign the time-stamped link to hyperlinks and define it as a way to represent the time dimension. This will show you how to incorporate temporal data into a variety of applications. These applications allow us to (1) monitor activity within a topical community as a function of time and (2) alter typical link-based ranking algorithms to capture timely authorities, which are on the rise today and should be rated higher than resources from previous days. They discussed a variety of elements for exploiting the time dimension in Web IR. They've proven their claims with a simple and easy-to-implement method that estimates the age of the page's content. Also, they suggested more rigorous processes for tracking temporal data, which would be useful for search engines that trawl the Web on a regular basis and keep a database of the content they find. The gathered temporal data can be utilized by search engines to create static link-based page rankings.

Liyanage et al. [9] created a more "intelligent web" and studied the majority of research focusing on data mining the web. The study of information retrieval (IR) has become one of the driving elements in achieving this goal. Clustering algorithms like K-means, Buckshot, Fractionation, and Suffix Tree Clustering are used in existing web IR systems. Many of these document clustering techniques rely on off-line clustering of the whole document collection. However, Web search engine collections are far too large and fluid for this to be feasible. As a response, clustering must be performed on the considerably lower number of documents returned as a result of a query. The CPU cycles and RAM given to each query are severely constrained since search engines provide free services to millions of consumers every day. Clustering is typically carried out on a separate system that takes search engine results, creates clusters, and shows them to the user.

Buntine et al. [10] proposed more advanced language models to be employed in information retrieval with some real success. It has been demonstrated experimentally that this is not always the best route to take. The following is an explanation of the issue: A generic statistical model built on a full global news corpus typically lacks the fidelity of these two key phrases when answering a query about "German immigration"; if possible, a statistical model focusing solely on "German immigration" would be ideal. The statistically based language modeling method for information retrieval still needs work. A different approach is taken to information retrieval by including statistical models, such as those created with discrete principal component analysis (PCA). This method is motivated by the widely held belief that, while people would like to be able to target their searches to specific locations, it is generally difficult to do so.

Several methods to address the challenge of content-based image retrieval on the Web have been presented by Voutsakis et al. [11]. Some of them are being implemented in research prototypes and commercial systems. The fourth group of systems includes general-purpose image search engines like Google Image Search. However, the research above, which focuses mostly on image and textual material, does not show how to choose high-quality Web sites and photographs connected to the query. HITS and other link analysis technologies assess the quality of Web sites as well as their relevance to the query's topic. Based on HITS, it shows how to handle sites that link to photos and pages that contain images. A set of 1,000 image samples is developed, including 500 logo-trademark images and 500 photographs of various categories. The feature vectors are used to train a decision tree to recognize logo and trademark images. The algorithm's predicted classification accuracy is 85%. The decision tree determines the likelihood of each image having a logo or trademark, known as “Logo-Trademark”.

Lewandowski [12] suggests the problems that search engines face when indexing the Web, as well as their solutions and user behavior. The primary goal of this study is to describe the shift from traditional information retrieval to Web information retrieval, which raises a number of new challenges. Search engines are the most popular implementation of information retrieval techniques into systems used by millions of people every day. Today's search engines rely heavily on link-based ranking algorithms, and it's easy to forget that these methods have drawbacks and might lead to bias in the results. These algorithms' most significant bias factors are described. They are built around a quality model, and quality is associated with power. Other quality indicators are ignored, and the algorithms are purely based on a citation indexing-based improved quality model. The rationale for this is mostly due to the Web's relatively simple link structure, the reliance on well-established bibliometric methods, and the reasonableness of the basic assumption.

Liu et al. [13] proposed a topology for graphs that are frequently formed in such a way that nodes are associated with data examples and edges show the relationship between examples in non-probabilistic models. Data examples in machine learning could be those having class labels or those to be assigned with class labels, grouped or rated. Those data examples always have concrete meanings when applied to an information retrieval task situation. To model the Web and user groups, one designs a network with nodes representing websites and edges signifying the existence of hyperlinks between them. Make a graph with nodes representing people and edges indicating hyperlinks between them.

Kolda et al. [14] proposed link analysis approaches that need to be improved. The internet's size continues to expand. TOPHITS stands for Topical HITS, which is a higher-order version of the well-known HITS paradigm. TOPHITS inserts anchor text information into a third dimension to produce an adjacency tensor. HITS and TOPHITS are two of numerous web-analysis approaches based on the adjacency matrix of a graph comprising a collection of online pages, similar to Page Rank and HITS. For a broad overview of these techniques, the primary eigenvector of a Markov matrix of page transition probabilities, a normalized version of the adjacency matrix, and a random-surfer component are used to calculate Page Rank scores. HITS, on the other hand, calculates both hub and authority ratings for each node, which correspond to the adjacency matrix's principal left and right singular vectors (though it can also be modified to include a type of random-surfer component).

De Lucia et al. [15] proposed the value of providing software developers with approaches and tools to assist with traceability recovery, which has become increasingly acknowledged in recent years. Several scholars have lately employed information retrieval (IR) techniques to recover traceable links between various artefact kinds. IR-based techniques recreate traceability linkages based on the similarity of the texts present in the software artefacts. The fact that most software documentation is text-based or gives textual descriptions, and programmers create source code IDs using valuable domain terminology justifies them. Based on a performance analysis conducted on software repositories of completed projects, they argue that IR approaches can assist the software engineer. They analyze the outcomes of IR approaches against a traceability matrix intended to contain the right relationships between the repository's artifacts for each experiment. They recommend using an incremental traceability recovery strategy to find a threshold that strikes a reasonable balance between accurately recovered links and false positives. This technique allows the software developer complete control over the threshold, allowing him or her to select the "best" threshold for each individual scenario.

The Internet raises a variety of new difficulties in terms of information retrieval while simultaneously providing new sources of knowledge. Link analysis has been presented as the primary method for automatically determining the quality of a page by Mandl et al. [16]. The major quality measure is the number of links pointing to a page. There have been many algorithms developed for link analysis, the most well-known of which is arguably the Page Rank algorithm. However, link analysis has significant flaws. Link assignment is a social process that results in astonishing worldwide patterns. A power law distribution governs the quantity of in-links on a web page. The median value is substantially lower than the average in such a distribution. Many pages have a limited number of in-links, but a few pages have a huge number. Link analysis algorithms' consequences Web search engine developers should take these results into account. Better link structure exploitation could lead to better retrieval results. Link-based metrics do not currently account for a page's hierarchical location. On the other hand, a deep in the hierarchy page with a large number of backlinks deserves to be given a higher authority value. Furthermore, the research demonstrates how web page authors believe catalogue pages are. They would rather have more options than be directed to a specific page.

Rather than targeting pages on lower hierarchical levels with a narrow topical focus, they assume that users will benefit from more browsing possibilities. The creators of web pages emphasize the relevance of surfing as a strategy for finding information in human-computer interaction.

Bazan et al. [17] present large networks, such as the Internet, geographic systems, transportation, and automatically generated social network databases, that require information management with a graph-like structure. In order to acquire more thorough information, users want not just basic tabular data from entities but also relationships with other entities via explicit or implicit values and links. Rather than a list of results, users frequently seek a collection of related entities that meet a certain condition. Graphs are the most natural approach to display findings in these instances. As a result, traditional database management systems (DBMS), which are frequently based on the relational model, may not be the ideal solution for responding to queries with these objectives. The user may be interested in not just recognizing a given author or publication but also evaluating relationships between authors, comprehending the significance of a specific work, or any other topic involving the research of interactions between objects. These conditions impose three fundamental problems: (i) the need for a versatile querying system that allows for IR queries of various flavors ranging from key word searches to complex graph mining; (ii) the need to integrate data from various sources to enrich the answers to complex queries over complete databases; and (iii) the need to integrate data from various sources to enrich the answers to complex queries over complete databases. Several graph database models (GDMs) and IR systems have been proposed to solve these challenges in part. In GDMs, data structures are described as graphs or extensions of graphs, and data manipulation is expressed using graph-oriented operations and type constructors.

Jin et al. [18] proposed a system to combine information retrieval and link-analysis methodologies over the extracted characteristics supplied by the Information Extraction (IE) engine to locate new knowledge. Here is the algorithm for building and ranking idea chains. First, the concept association graph (CAG) may undergo an optional cleaning procedure, depending on the demands of the user. As filters, support and mutual information (edge weight) are employed. High mutual information item sets, as well as high support item sets, are almost always significant. The support and edge confidence parameters can be used to change the size of the intended CAG (which will filter the edges whose weights are below a user-specified threshold).

As a result, estimating the authority of individuals who publish answers on such quality assurance (QA) portals without depending just on user input is becoming increasingly critical, as presented by Jurczyk et al. [19]. While assessing author authority in collaborative portals such as Yahoo! Answers has been a popular focus of research, the problem remains unsolved. Previously, numerous characteristics such as author activity, number of clicks on answers, and average length of postings were examined in order to choose the best responses for a certain question. They, on the other hand, focus on determining user authority, which may be used to categorize responses, find "experts," create incentive systems, and detect spam. They offer the HITS link analysis algorithm. The HITS method forecasts the significance of websites by assigning each page a hub and an authority value. A page is regarded as a good hub if it connects to authoritative pages, and authoritative pages are connected by excellent hubs. QA portals are intuitively comparable to this notion. Writers of questions might be regarded as hubs, and authors of responses as authorities.

Egozi et al. [20] introduced information retrieval (IR) systems that generally encode queries and documents using the Bag of Words (BOW) representation and retrieve results by looking for occurrences of query terms in indexed documents. Such methods are unable to find relevant materials that do not explicitly mention query terms, especially when users make very brief queries, as in Web searches. Explicit Semantic Analysis (ESA), a novel IR methodology based on the recently announced ESA, uses significant encyclopedic information to improve the basic BOW text representation with conceptual and semantically rich elements. A unique unsupervised learning technique that filters created features to meet the specific query corpus context, using the BOW results as examples, is also proposed to lessen the noise introduced by generated features (especially damaging to IR).

Singh et al. [21] discuss Web sites that are semi-structured, and Web content has a wide range of meanings, degrees of quality of information collected, and inferences formed from the extracted data via the World Wide Web (WWW). They are seeking more effective information retrieval strategies and technologies to help them locate, extract, filter, and retrieve the information they need. A search engine has three fundamental components. The three are the crawler, indexer, and ranking mechanism. A crawler is an internet-navigating robot or spider that downloads web pages. An indexing module receives the downloaded pages and parses and produces an index based on the keywords detected on them. The keywords are usually used to create an alphabetical index. When a user makes a query using keywords on a search engine's interface, the query processor component matches the query keywords with the index and delivers the URLs of the pages to the user. Search engines, on the other hand, utilize a ranking algorithm to present the most relevant websites at the top and the least relevant ones at the bottom before showing the sites to the visitor. It makes it easier for users to navigate search results. The Google search engine is immensely famous due to its Page Rank algorithm. Page ranking algorithms and web mining techniques are used by search engines to display search results based on relevance, significance, and content score, as well as user interest. Some ranking algorithms are simply dependent on the link structure of the document. Others search the documents for real content to establish their popularity scores (web structure mining) or web content mining.

Liu et al. [22] first mentioned in their dissertation that data mining is the process of obtaining information from a large database. A well-known data mining approach is information retrieval. Search engines are essential for finding information on the Internet and in other directories. The quality of a search engine is determined by the ranking functions that are utilized to offer results based on the user's query. Many information retrieval systems rely on ranking. Crawlers, sometimes known as robots, are three key processes in information retrieval. They automatically download web pages. The indexer is a keyword-based indexer. The indexer will get the document by matching the keywords to the user's query. A number of different ranking algorithms for document ranking based on queries have been proposed. He proposed a ranking system that was semi-supervised. It learns to rank material that is both labeled and unlabeled. A Support Vector Machine (SVM) was used to extract photos from the web. The method's fundamental idea is to combine text or visual attributes for automatic image ranking using query image systems.

Hussein et al. [23] compared and contrasted various ranking algorithms for ranking various websites, including Page Rank, Weighted Page Rank, and Hypertext Induced Topic Selection (HITS). Then, to improve search results, they proposed a Topic-Sensitive Page Rank and a Weighted Page Rank-based ranking algorithm. They used simulation tools to test the performance of existing algorithms and the suggested method in the experiment. They

demonstrated that when the proposed algorithm is employed instead of these techniques, the ranking of web pages in search results improves and becomes more accurate.

Gu et al. [24] used a web dataset that presents time graph patterns to forecast the temporal properties of the web. This dataset was generated by gathering authoritative blogs that are sensitive to hot topics, sites that summarize such blogs and news sites, as well as blogs with a high number of postings connected to daily news. Those from the sites, as well as pages that link to or from them, are collected. A web graph is received with 732 links and 850 pages split over 55 sites. According to the research, these time graph patterns can be utilized to examine both the structural and temporal features of the web.

Kumar, et al. [25] proposed Web mining with the following challenges: Web pages are semi-structured, and web data has a broad range of interpretations. Web mining methods, as well as techniques from other domains like databases (DB), information retrieval (IR), natural language processing (NLP), and machine learning (ML), can be used to address the difficulties listed above. Web mining is the technique of employing data mining technologies to automatically identify and extract information from the internet (WWW). Web structure mining helps users find relevant resources by researching the Web's link structure. NLP, ML, and other approaches can be used to overcome the problems mentioned above. Web mining is the technique of employing data mining technologies to automatically identify and extract information from the internet (WWW). Web structure mining helps users find relevant resources by researching the Web's link structure. The WWW may be represented as a directed graph G (V, E), where V stands for pages and E for links.

Oliveto et al. [26] proposed Maintaining dependencies between different types of software artifacts is widely recognized as an important support activity both during initial system development and also during the ongoing change management process Traceability links between the free text documentation associated with the development and maintenance cycle of a software system and its source code are helpful in a number of tasks such as requirement coverage, program comprehension, and impact analysis. They propose IR-based methods for a list of candidate traceability links on the basis of the similarity between the texts contained in the software artifacts. These approaches are based on the idea that two artifacts with high textual similarity share several ideas, making them good tracing candidates. Several IR techniques have been developed for traceability recovery, including vector space and probabilistic models, as well as Latent Semantic Indexing (LSI). Two measures are used to assess the retrieval accuracy of IR-based traceability recovery methods: recall and precision. The proportion of correct connections detected is measured by recall, whereas the percentage of correct links is measured by precision.

Bhatia et al. [27] used a range of rating algorithms based on online content mining and structure mining to examine the links that led to the rated pages. The three algorithms are Page Rank, Weighted Page Rank, and Weighted Page Content Rank. The following are the differences between these algorithms, according to them: The Weighted Page Rank reflects the most important web pages based on a weight value calculated from the number of inbound links to a page and all of the page's reference sites. The Weighted Page Content Rank, on the other hand, is based on web structure and online content mining techniques, which indicate the importance of a page and determine its relevance. After that, they compared which of these algorithms selected the best and most relevant pages for a specific query. They demonstrated that the Weighted Page Content Rank returns the best relevant page for a query.

Li et al. [28] proposed information retrieval (IR) systems are designed to find relevant information to meet the present user's information needs. To improve retrieval efficacy, other strategies can be used, such as personalized information retrieval (PIR), combining recommender systems (RSs) with IR, or exploiting inter-document link architectures via algorithms like Page Rank. Each of these methods has its drawbacks. PIR systems gather both explicit and implicit feedback to create a user profile with the goal of providing retrieval results that better fit the needs of individual users. However, in many cases, there may be no way to gather appropriate feedback information to help with the present inquiry, which is a substantial challenge for PIR systems. They conclude that employing either RSs or a Page Rank method can improve IR results based on their findings and analyses. In this paper, they offer a revolutionary way to improve standard IR by combining RSs, Page Rank, and IR. Recommender approaches are used to extract the correlation between pages in the dataset and then use the Page Rank algorithm to determine the relevance of each document based on this correlation.

Shaalan et al. [29] advocated in their study that Arabic information retrieval be improved. The process of retrieving all relevant documents from unstructured textual content in answer to a query is known as Information Retrieval (IR). Each document is represented by a set of keywords known as index terms in the traditional IR paradigm. Because of so many synonyms, the Arabic language has a huge vocabulary, which is a hurdle in the IR process. This research investigates query expansion using the Expectation Maximization algorithm (EM) to enhance the number of relevant documents retrieved. It also looks at the best EM distance for Arabic words to see how similar they are. The key contributions of this research are Improving Arabic Information Retrieval by expanding Arabic queries with comparable index phrases and Query expansion based on similarity of terms for improving Arabic IR 175.

Akila et al. [30] advocated clustering linked data and documents in order to obtain helpful information for information retrieval systems. There have been various approaches to dealing with clustering. The most important of these is K-Means, which is used to divide a data set into several homogenized groups based on their similarities. K-Means results are frequently optimized using metaheuristic techniques. The Honey Bee Algorithm, as a Meta-Heuristic Algorithm, presents a fresh technique to solution space search by monitoring honey bee foraging behavior, which improves solution quality. Intelligent approaches are mostly employed in the development of outstanding current and professional information systems for the solution of complex problems and to make difficult sets of data or information better.

Chiu et al. [31] recommend emerging technology as a supplement or alternative solution, as well as a way to launch a new firm. It is also crucial for a company or industry to understand where a possible technology started so that the firm can decide whether or not to invest resources in it and the industry can watch the technology's progress. Researchers and practitioners can delve into the patent database to evaluate if a developing technology has any promise. A rare information retrieval methodology will be proposed to identify uncommon patents from the patent database, as well as a link strength measure method to separate noteworthy rare patents from the possible ones. Cluster Analysis is designed to construct thin-film solar cell clusters by performing cluster analysis utilizing two-step clustering and cluster identification. In order to do cluster analysis, two-step clustering from SPSS Clementine is used to group patents into groups. For large data sets, two-step clustering is a scalable cluster analysis method. It has the ability to handle either continuous or categorical data (or attributes). There is only one data pass required.

Thwe et al. [32] proposed several attempts that have been made to take advantage of web page access prediction by preprocessing web server log data and analyzing web users' browsing behaviors. The idea is to figure out how to use data from internet logs to forecast how people will access websites. Because of its high precision, the Markov model is the most extensively used in pattern recognition and prediction models based on the sequence of previously read pages. They are strong candidates for sequential pattern discovery for link prediction due to their suitability for mimicking sequential processes. The most extensively used link analysis approach is Page Rate, which is used to rate the search engine results delivered after a user query. The value of a page is assessed in relation to its connectivity to and from other relevant pages to establish the grade.

Dit et al. [33] suggested software systems for bug fixes, speed and stability improvements, and new feature additions. Feature or idea localization, which is the activity of finding the source code that implements functionality, also known as a feature, is a crucial aspect of the program understanding process. Before making changes to a feature, software developers must first locate and comprehend its implementation. Feature location can be a time-consuming operation for software engineers who are unfamiliar with a system. Data fusion is the concept of combining data from several sources. Individual data sources have advantages and limits, but when they are integrated, those disadvantages are minimized and better results are achieved. A call graph is another way to represent software in graph form. Methods are represented by nodes, while links or calls between methods are represented by edges. As a result, web mining techniques can be used to identify relevant information from software's structure, such as crucial classes for program comprehension, component ranks in software repositories, and assertions that can be refined from idea bindings. This research investigates if web mining may be used to locate features, either as an independent tool or as a filter to an existing strategy.

Dit et al. [34] defined Web mining as the process of extracting knowledge from Web data using data mining techniques, with at least one type of structural (hyperlink) or usage (Web log) data used in the process (along with or without other types of Web data). Web mining duties are divided into three categories: web content mining, web structure mining, and web use mining. Online structure mining is a technique for determining the structure of a website or web page. The goal of web structure mining is to figure out how hyperlinks are linked between documents. Web structure mining will categorize Web pages and create information such as website similarities and linkages. In the vast majority of web data retrievals, the HITS algorithm finds the hubs and authorities in a community on a specific topic or query. In research on hyperlinked document link analysis, HITS was used for the topic of topic distillation, and several forms of link weights were utilized to reflect the relevance of links in hyperlinked documents. As a result, link analysis is essential for finding related material based on hypertext and external evidence, and it improves performance greatly. The anchor text is employed as an external source of evidence in the WHITs algorithm, and it outperforms for short-term enquiries.

Prasanth et al. [35] present Web Usage Mining (WUM) as a project that examines web log files using a method for uncovering knowledge in large databases. Actually, websites generate a large amount of web log data, which contains information about the web user's behavior. Analysis and discovery of useful information from the WWW is what the WUM calls it. It is also known as "the application of data mining technologies to massive web data repositories," or "automatically recognizing web user access patterns from various types of web servers." The web suggestion system is essential to making user web page access easier. The WUM technique can be used to anticipate web page access in the future. The suggested system predicts user

navigational preferences based on previous behaviors, and the learned pattern is then used to recommend more appropriate websites to users. Before recommending the pages to the user, a collaborative filtering process was used to compare a user's previous visiting preferences with those of other users with similar interests.

Irfan et al. [36] combined diverse methodologies and technologies; they were able to view the information from various perspectives, allowing them to locate valuable information. The World Wide Web has grown into a massive and popular library of hyperlinked papers carrying vast amounts of information that assist in meeting the wants and requirements of all humans. The web is a data store that houses the bulk of documents in HTML format. The World Wide Web is rapidly developing, and with it comes a plethora of data mining techniques that aid in the discovery of hidden information and important trends. This concealed data assists the user with a variety of activities, including taking user feedback, making decisions, customer relationship management, and website reorganization. Artificial intelligence, natural language processing, and other statistical approaches are combined with mining methods to improve the efficiency of mining tasks. Web mining collects valuable information from websites using a variety of tools and algorithms that work with semi-structured and unstructured data. Online content mining focuses on the content of websites to obtain information, whereas web structure mining keeps track of links between pages and serves the most relevant pages to the user based on their query. Web use mining considers web log information on web pages, which is useful for tracking user data. K-Nearest-Neighbor is based on training by analogy, which compares training and test tuples. Each of the training tuples represents a point in the n-dimensional pattern space and has n attributes. An n-dimensional pattern space is used to store all of the training tuples. In pattern space, the K-Nearest-Neighbor classifier seeks the tuple that is closest to an unknown tuple. K-Nearest-Neighbor classifiers may also be used to forecast an unknown tuple because they produce a real-valued prediction. Because each characteristic is assigned the same weight when utilized for distance-based comparisons, it has low accuracy for noise and irrelevant features.

Janani et al. [37] analyze online links and extract the relevant information from a collection of web articles. Two pattern matching algorithms were given. The Boyer-Moore algorithm (Turbo BM) and Backward Nondeterministic Dawg Matching (BNDM) are the algorithms in question. They evaluated the models' performance using the following metrics: time, iteration count, and accuracy results. In terms of the three above measures, they found that the (BNDM) algorithm performed better than the Turbo BM technique.

Denture et al. [38] proposed data mining as an interdisciplinary topic concerned with collecting information from large amounts of data and putting it into a structure that can be easily analyzed for later use. Document Information Retrieval (DIR), the first IR issue researched, is one of them. Using three different algorithms (k-means, DBSCAN, and Spectral), the item database is separated into clusters, with similar things grouped together. The purpose is to limit the number of phrases that are common between object clusters. Two transformation techniques (boolean and weighted) are presented to adjust the pattern mining algorithms in looking for significant patterns in each cluster of objects. The Boolean technique transforms a collection of objects into a transaction database without considering the frequency of the objects' terms, whereas weighted techniques consider the objects' frequency throughout the transformation process. Two pattern mining approaches are utilized to find significant patterns for each cluster of objects.

3. Related Work Summary

The survey summarizes some prior work on link analysis in web information retrieval using pattern mining. Table 1 summarizes past work by year of completion, key aims, algorithms utilized, and results.

Table 1: Previous Research Papers Summarization

Ref	Year	Objectives	Algorithm	Results
[5]	1993	The goal of this algorithm is to generate an interior icon arrangement that ensures no two icons are in the same place. The principle of rank layout is what we call it.	1) Create a layout method that allows us to create Info Crystals with N inputs. 2) It's known as the rank layout principle since it firmly adheres to the rank coding theory.	Info Crystal elements can be selectively visualized to emphasize qualitative or quantitative information. Graphically specifying vector-space queries. Users can develop and maintain complicated search queries using Info Crystal in conjunction with a query outlining and navigation tool.
[6]	1994	It is to simplify a hypertext's structure into a hierarchy and a set of cross-reference connections.	Navigational Algorithm: the “Fish-Search”.	Finding the "best" order for retrieving nodes from an algorithm. It transforms the browser into a World Wide Web information retrieval tool.
[2]	2000	They analyzed hyperlinks in web information retrieval.	1) Query Independent Connectivity Based Ranking. 2) Query Dependent Connectivity Based Ranking.	These algorithms are useful for analyzing hyperlinks in web information retrieval. Five link-based metrics (in-degree, out-degree, HITS authority score, HITS hub score, and PageRank) were used to evaluate the quality of an information retrieval system.
[7]	2004	When modeling, a hyperlink graph contains useful information.	1) Kleinberg's (HITS) algorithm improves on the basic concepts of hubs and authority. 2) To provide a detailed examination of the HITS algorithm, emphasizing the importance of in and out degrees.	HITS assigns hubs and authority relevance scores and calculates them in a mutually reinforcing manner. Two link-based metrics (HITS authority score and HITS hub score) were used to evaluate the quality of the system.
[8]	2004	Search engines track and save temporal data for each page in their repository. Over time, track the level of activity in a relevant community. Traditional link-based ranking systems to capture timely authorities.	There are several elements to exploiting the time dimension in the context of Web IR.	More sophisticated processes for tracking temporal data were proposed, which are suitable for search engines that continuously browse the Web and retain a repository of the resources found.
[9]	2004	To create a more "intelligent web".	Clustering algorithms such as K-means and other clustering techniques.	Clustering is frequently done on a separate system that takes in search engine results, forms clusters, and displays them to the user.
[10]	2005	Information retrieval is beginning to use complex language models.	Using statistical models.	A generic statistical model constructed on a full worldwide news corpus typically lacks fidelity on these two key

				phrases together when answering a query regarding "German immigration." If it were possible, a more particular statistical model about "German immigration" would be ideal. A link-based metric (Topic specific scoring) was used to evaluate the quality of information
[11]	2005	Some have been implemented in research prototypes and commercial systems for content-based picture retrieval on the Web.	HITS and other link analysis methods.	Estimate the quality of the Web pages as well as their topic relevancy to the query. This demonstrates how to deal with pages that contain photos and pages that link to images. The method is represented by a precision-recall curve. A method is better than another if it achieves better precision and recall.
[12]	2005	This demonstrates the transition from traditional to web-based information retrieval.	In today's search engines, link-based ranking algorithms are the most common.	The algorithms are purely based on a citation indexing-based improved quality model. The rationale for this is mostly due to the Web's relatively simple link structure, the reliance on well-established bibliometric methods, and the reasonableness of the basic assumption. The quality of information retrieval systems in general and search engines in particular is measured with standard measures like recall and precision.
[13]	2006	To form graph topology in such a way that nodes are associated with data examples and edges show the relationship between examples in non-probabilistic models and data to be clustered and ranked.	1) Semi-supervised Learning 2) Concerned with the training of both labeled and unlabeled data. 3) The assumption that similar data examples are consistent.	Only a few algorithms have unambiguous graph interpretations or could be connected to graphs. Five link-based metrics (traffic rank, manifold rank, HITS authority score, HITS hub score, and PageRank) were used to evaluate the quality of the system.
[14]	2006	Suggested link analysis approaches that need to be improved.	1) PageRank. 2) HITS analyzing methods.	The entries of the primary eigenvector of a Markov matrix of page transition probabilities determine PageRank scores. HITS calculates each node's hub and authority scores.
[15]	2006	Information retrieval (IR) techniques were used to solve the problem of retrieving traceability relationships between different types of artefacts.	IR-based techniques recreate traceability linkages based on the similarity of the texts present in the software artefacts. IR-based	They suggest an incremental traceability recovery approach with the goal of determining a threshold that achieves a good balance between correctly

			techniques recreate traceability linkages based on the similarity of the texts present in the software artefacts.	retrieved connections and false positives. Rather than recovering all correct links, it might be more effective to identify an "optimal" threshold that achieves a good balance between precision and recall.
[16]	2007	Enabling information retrieval while also providing extra knowledge sources.	The PageRank algorithm, as well as link analysis techniques, have various flaws.	Link analysis has been used to assess the quality of a page. A deep in the hierarchy page with a large number of backlinks deserves to be given a higher authority value. Web page authors stress the importance of navigation.
[17]	2007	The necessity to organize information having a graph-like character has arisen. Users are interested in getting not only plain tabular data from entities but also relationships with other entities that leverage explicit or implicit values and linkages to gain more detailed information.	There have been several Graph Database Models (GDMs) and IR systems proposed.	GDMs are those in which the schema and instances' data structures are modeled as graphs or generalizations of graphs, and data manipulation is defined using graph-oriented operations and type constructors.
[18]	2007	An information extraction (IE) engine is used to find new knowledge.	Depending on the needs of the user, the concept association graph (CAG) may go through an optional cleaning process.	Support and mutual information (edge weight) are used as filters. High support item sets, as well as high mutual information item sets, are virtually always significant. The evaluation looked at (i) whether the target chains were found among the top 5 choices by the retrieval models; (ii) the average rank of the target chains; and iii) the average recall achieved by both models at various lengths.
[19]	2007	Without relying just on user feedback, it is becoming increasingly vital to automatically measure the authority of people that publish responses on such QA portals. While evaluating the authority of writers in collaborative portals has been a hot topic in the scholarly community,	The HITS method was created to forecast the relevance of web pages by allocating a hub and an authority value to each page.	If a page links to authoritative pages, it is considered a good hub, and authoritative pages are linked by good hubs. This concept is intuitively similar to QA portals. Evaluate the methods by comparing our estimated scores with measures obtained derived from explicit user feedback. In fact, votes and stars scores are instances of general feedback mechanism design, where the former allows "popularity" feedback from all users, whereas the latter supports "quality" feedback from the user posing the original question.

[20]	2008	When retrieving results, identify occurrences of query phrases in indexed documents.	1) An innovative unsupervised learning algorithm. 2) Explicit Semantic Analysis (ESA).	ESA enhances the conventional BOW text representation with conceptual and semantically rich features using a first algorithm that filters the generated features to match the specific query corpus context. IR performance is measured by Mean Average Precision (MAP).
[21]	2009	Are you seeking more effective information retrieval techniques and technologies to help you locate, extract, filter, and retrieve the information you need?	1) Crawling, indexing, and ranking mechanisms 2) PageRank algorithm.	A crawler is a spider that crawls the internet and downloads pages. The pages are downloaded and passed to an indexing module, which parses them and creates an index based on the keywords. PageRank uses web mining algorithms to offer search results based on relevance, significance, and content score. A link-based metric (PageRank score) was used to evaluate the quality of the information retrieval system.
[22]	2009	The quality of a search engine is determined by the ranking functions which are used to provide the results according to a user's query.	1) Crawler and indexer operations 2) PageRank algorithm.	Crawlers, sometimes known as robots, download online pages automatically. The Indexer is a keyword-based indexer. Ranking is at the heart of many information retrieval systems, and the indexer will match the keywords depending on the user's query and return the content. Several evaluation measures have been proposed and used in IR systems, such as mean reciprocal rank (MRR), mean average precision (MAP), discounted cumulative gain (DCG), and rank correlation (RC).
[23]	2010	Web page ranking.	1) PageRank. 2) Weighted PageRank. 3) HITS. 4) The proposed algorithm.	The proposed algorithm gave more accurate search results. Evaluate the ranking algorithms over a range of accepted information retrieval metrics, namely Precision at K (P (K)) and Normalized Discounted Cumulative Gain (NDCG).
[24]	2010	Analyze the structural and temporal aspects of the web.	Time graph patterns.	The analysis was done effectively using these patterns.
[25]	2010	Web mining is the process of automatically discovering and extracting information from the internet using data mining tools (WWW). Web structure mining analyzes the link structure of the Web to	(NLP), machine learning.	Web mining techniques, as well as other fields such as database (DB), information retrieval (IR), natural language processing (NLP), and machine learning, have a wide range of meanings, degrees of

		assist users in finding relevant documents.		information quality, and conclusions drawn from the information mined. Five link-based metrics (in-degree, out-degree, HITS authority score, HITS hub score, and PageRank) were used to evaluate the quality of the system.
[26]	2010	Maintaining dependencies between various types of software artifacts.	Latent Semantic Indexing (LSI), or vector space, and probabilistic models.	The retrieval accuracy of IR-based traceability recovery methods is measured using two metrics: recall and precision. Recall measures the percentage of accurate links identified, while precision measures the percentage of right links found.
[27]	2011	To analyze the hyperlinks that lead to the ranked pages.	1) Weighted Page Rank. 2) Page Rank. 3) Weighted Page Content Rank.	The weighted Page Content Rank gives the most relevant page for a given query. A link-based metric (PageRank score) was used to evaluate the quality of the system.
[28]	2011	Identify appropriate information to meet the information needs of the current user. A variety of strategies can be used to improve retrieval efficacy.	IR and PageRank.	We use recommender approaches to extract the correlation between texts in the dataset and then use the PageRank algorithm to calculate each document's relevance. The Mean Average Precision (MAP) was used to evaluate overall retrieval effectiveness and precision at cut-off n to evaluate how early the relevant documents were retrieved.
[29]	2012	Expanding Arabic queries with similar index phrases to improve Arabic information retrieval. Improving Arabic IR 175 by Query Expansion Based on Similarity of Terms.	The EM algorithm is being used in research to improve the quantity of relevant documents returned.	The EM distance is a crucial aspect of this system's overall performance. EM eliminates the system's needless retrievals of answers based on dissimilar keywords. The Arabic information retrieval system's structure. It is divided into two stages: indexing and querying. The Precision-Recall and the Precision at document cutoff n were used to prove that expanding queries retrieve more relevant documents.
[30]	2012	Obtain vital information for information retrieval systems.	K-Means results are frequently optimized using meta-heuristic techniques. As a meta-heuristic algorithm, the Honey Bee algorithm.	Introduces an innovative approach to solution space search based on honey bee foraging behavior, which leads to improved solution quality. Intelligent approaches are mostly used in the development of outstanding current and professional information

				systems for resolving complex challenges.
[31]	2012	Examine the patent database for new technological possibilities. A rare information retrieval approach will be developed to discover unusual patents in the patent database.	"Cluster Analysis" is a program that uses two-step clustering and cluster identification to do cluster analysis.	For large data sets, two-step clustering is a scalable cluster analysis method. It has the ability to handle both continuous and categorical data (and attributes). There is only one data pass required.
[32]	2013	Predicting web page access by evaluating web users' browsing patterns and preparing web server log files.	1) To identify patterns, the Markov model is used. 2) Page Rank is the most commonly used link analysis algorithm.	Model of Markov pattern recognition using the sequence of previously accessed pages and a prediction model with excellent accuracy. Page Rank is used in order to rank the results supplied by a search engine after a user query.
[33]	2013	Software systems to fix bugs, improve performance and stability, and add new features.	Web mining algorithms are easily transferable to software.	Discover important information from its structure, such as crucial classes for program comprehension, component ranks in software repositories, and refined assertions from concept bindings. The effectiveness measure is an accepted metric to evaluate feature location techniques.
[34]	2013	The mining technique uses either structured (hyperlink) or usage (Web log) data to derive knowledge from Web data.	Based on mutual reinforcement, the HITS (Hyperlink Induced Topic Search) algorithm.	Provides a novel approach for searching the web and distilling themes. In hyperlinked papers, to emphasize the importance of links. As a result, link analysis is critical for detecting related information utilizing hypertext and external evidence, and it significantly enhances performance.
[35]	2017	A web recommendation system is necessary to facilitate user web page access.	The WUM technique can be used to forecast future web page access.	The WUM technique can be used to anticipate web page access in the future. The suggested system predicts user navigational preferences based on previous behaviors, and the learned pattern is then used to recommend more appropriate websites to users. Precision, recall, F Measure, and response time were used to evaluate the quality of the system.
[36]	2018	With the increase in data, several mining approaches are being utilized to aid in the discovery of buried data and important patterns. This concealed data assists the user with a variety of activities, including taking	Artificial intelligence is being used with mining method.	Improve the efficiency of mining tasks. Online content mining focuses on the content of websites to obtain information, whereas web structure mining keeps track of links between pages and displays the most relevant

		user feedback, making decisions, customer relationship management, and website reorganization.		pages to the user based on the query.
[37]	2019	Retrieve information from web links.	1) Turbo BM algorithm. 2) Backward Nondeterministic Dawg Matching (BNDM) algorithm.	The accuracy of BNDM is higher than Turbo BM. Search time, number of iterations, and relevancy were used to evaluate the quality of the system.
[38]	2021	Information retrieval is the process of extracting meaningful information from a collection of elements (IR). There have been several types of IR concerns discussed.	Three unique algorithms (k-means, DBSCAN, and Spectral).	Divide the item database into clusters, with similar items grouped together. The purpose is to limit the number of phrases that are common between object clusters. To evaluate the retrieved objects, the mean average precision (MAP) and the F measure were used.

This paper presents an examination of many related studies and algorithms utilized in the ranking of online sites. It discusses the advantages and limitations of these algorithms. The analysis of the obtained results shows clearly the mean of performance. There is a need to develop a new approach relevant for computing the score of a web page and to explore several ways in order to improve the different algorithms proposed by the authors in a related study.

The word sense disambiguation must be done in order to comprehend what a user is looking for. When a term is ambiguous, meaning it can have several meanings, the disambiguation process is started, thanks to which the most probable meaning is chosen from all those possible. Such processes make use of other information present in a semantic analysis system and take into account the meanings of other words present in the sentence and in the rest of the text.

It is therefore better to use a semantic search engine to increase search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data area, whether on the Web or within a closed system.

4. Conclusions

This survey paper aims to review and analyze several related research papers that are related to link analysis in web information retrieval using various algorithms based on several different aspects: the research year, the research objectives, the algorithms used to complete their research, and the results obtained after applying the algorithms. These results revealed that assessing hyperlinks in web information retrieval is the most important goal of these research publications. Researchers also employ the page rank, weighted page rank, and weighted page content rank algorithms to properly analyze hyperlinks in web information retrieval.

References

[1] M. Bhatia and A. Kumar, "Paradigm shifts: From pre-web information systems to recent web-based contextual information retrieval," *Webology*, vol. 7, no. 1, pp. 1–7, 2010.

[2] D. B. Lomet *et al.*, "Editorial Board Editor-in-Chief TC Executive Committee," vol. 23, no. 3, 2000, [Online]. Available: http://www.research.microsoft.com/research/db/debull.

[3] R. Jain, A. Chavan, and S. Nair, "Web Structure Mining using Link Analysis Algorithms," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 4969–4973, 2015.

[4] G. Buehrer and K. Chellapilla, "A scalable pattern mining approach to web graph compression

with communities," *WSDM'08 - Proc. 2008 Int. Conf. Web Search Data Min.*, pp. 95–106, 2008, doi: 10.1145/1341531.1341547.

[5] A. Spoerri, "InfoCrystal: A visual tool for information retrieval," *Proc. 4th Conf. Vis. VIS 1993*, no. November 1993, pp. 150–157, 1993, doi: 10.1109/visual.1993.398863.

[6] P. de Bra, G.-J. Houben, Y. Kornatzky, and R. Post, "Information Retrieval in Distributed Hypertexts," *Riao*, no. September 2014, 1994.

[7] C. H. Q. Dingt, H. Zha, X. He, P. Husbands, and H. D. Simon, "Link analysis: Hubs and authorities on the world wide web," *SIAM Rev.*, vol. 46, no. 2, pp. 256–268, 2004, doi: 10.1137/S0036144501389218.

[8] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer, "Trend detection through temporal link analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 14, pp. 1270–1281, 2004, doi: 10.1002/asi.20082.

[9] H. Liyanage and G. E. M. D. C. Bandara, "Macro-clustering: Improved information retrieval using fuzzy logic," *IEEE Int. Symp. Intell. Control - Proc.*, no. October, pp. 413–418, 2004, doi: 10.1109/isic.2004.1387719.

[10] W. Buntine, J. Löfström, S. Perttu, and K. Valtonen, "Topic-specific link analysis using independent components for information retrieval," *AAAI Work. - Tech. Rep.*, vol. WS-05-07, pp. 47–52, 2005.

[11] E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Weighted link analysis for logo and trademark image retrieval on the Web," *Proc. - 2005 IEEE/WIC/ACM Int. Web Intell. WI 2005*, vol. 2005, pp. 581–585, 2005, doi: 10.1109/WI.2005.162.

[12] D. Lewandowski, "Web searching, search engines and Information Retrieval," *Inf. Serv. Use*, vol. 25, no. 3–4, pp. 137–147, 2005, doi: 10.3233/isu-2005-253-402.

[13] Y. Liu, "Graph-based learning models for information retrieval: A survey," 2006, [Online]. Available: http://www.cse.msu.edu/~rongjin/semisupervised/graph.pdf

[14] T. Kolda and B. Bader, "The TOPHITS model for higher-order web link analysis," *Work. Link Anal. Counterterrorism Secur.*, vol. 7, pp. 26–29, 2006, [Online]. Available: papers://939f4cb1-7db6-4526-91fa-0f6dbb0b8d74/Paper/p3402

[15] A. De Lucia, F. Fasano, R. Oliveto, and G. Tortora, "Can information retrieval techniques effectively support traceability link recovery?," *IEEE Int. Conf. Progr. Compr.*, vol. 2006, no. September 2015, pp. 307–316, 2006, doi: 10.1109/ICPC.2006.15.

[16] T. Mandl, "Link Analysis and Site Structure: Refining Web Information Retrieval," pp. 1–6, 2007, [Online]. Available: http://www.uni-hildesheim.de/~mandl/Publikationen/webir_ab.pdf

[17] N. Martínez-Bazan, J. Nin, V. Muntés-Mulero, M. A. Sánchez-Martínez, S. Gómez-Villamor, and J. L. Larriba-Pey, "DEX: High-performance exploration on large graphs for information retrieval," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. July 2014, pp. 573–582, 2007, doi: 10.1145/1321440.1321521.

[18] W. Jin, R. K. Srihari, H. H. Ho, and X. Wu, "Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, no. November 2007, pp. 193–202, 2007, doi: 10.1109/ICDM.2007.62.

[19] P. Jurczyk and E. Agichtein, "HITS on question answer portals: Exploration of link analysis for author ranking," *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR'07*, pp. 845–846, 2007, doi: 10.1145/1277741.1277938.

[20] O. Egozi, E. Gabrilovich, and S. Markovitch, "Concept-based feature generation and selection for information retrieval," *Proc. Natl. Conf. Artif. Intell.*, vol. 2, pp. 1132–1137, 2008.

[21] A. K. Singh and R. K. P, "A Comparative Study of Page Ranking Algorithms for Information Retrieval," *Comput. Eng.*, vol. 3, no. 4, pp. 469–480, 2009.

[22] T. Y. Liu, "Learning to rank for Information Retrieval," *Foundations and Trends® in Information Retrieval*, Vol. 3: No. 3, pp 225-331, 2009. doi: 10.1561/1500000016.

[23] M. Hussein and M. Mousa, "An Effective Web Mining Algorithm using Link Analysis," *International Journal of Computer Science and Information Technologies*, vol. 1, no. 3, pp. 190–197, 2010.

[24] Y. Gu et al., "Detecting hot events from web search logs," Web-Age Information Management. WAIM 2010. Lecture Notes in Computer Science, vol 6184. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14246-8_41

[25] Kumar, "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval," *Am. J. Appl. Sci.*, vol. 7, no. 6, pp. 840–845, 2010, doi: 10.3844/ajassp.2010.840.845.

[26] R. Oliveto, M. Gethers, D. Poshyvanyk, and A. De Lucia, "On the equivalence of information retrieval methods for automated traceability link recovery," *IEEE Int. Conf. Progr. Compr.*, pp. 68–71, 2010, doi: 10.1109/ICPC.2010.20.

[27] T. Bhatia, "Link analysis algorithms for web mining," *Int. J. Comput. Sci. Telecommun. Vol. 2*, vol. 4333, no. 2, pp. 243–246, 2011.

[28] W. Li, "Exploring Recommenders for Improved Information Retrieval," in Fourth BCS-IRSG Symposium on Future Directions in Information Access (FDIA 2011), September, 2011, doi: 10.14236/ewic/fdia2011.7.

[29] K. Shaalan, S. Al-sheikh, and F. Oroumchian, "Expansion Based-on Similarity of TermsQuery," *IFIP Int. Fed. Inf. Process. 2012*, pp. 167–176, 2012.

[30] D. Akila, C. Jayakumar, G. Shree, and S. Jain, "Link-based Ensemble Approach for Web Information Retrieval Using Honey Bee and K-Means Algorithm," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 11, pp. 75–81, 2012, [Online]. Available: www.ijarcsse.com

[31] T. F. Chiu, C. F. Hong, and Y. T. Chiu, "Emerging technology exploration using rare information retrieval and link analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7654 LNAI, no. PART 2, pp. 540–549, 2012, doi: 10.1007/978-3-642-34707-8_55.

[32] P. Thwe, "Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm," *Int. J. Sci. Technol. Res. Vol. Ijstr*, vol. 2, no. 3, pp. 240–246, 2013.

[33] B. Dit, M. Revelle, and D. Poshyvanyk, "*Integrating information retrieval, execution and link analysis algorithms to improve feature location in software*," Empir *Software Eng*, vol. 18, no. 2. 2013. doi: 10.1007/s10664-011-9194-4.

[34] B. Dit, M. Revelle, and D. Poshyvanyk, "Integrating information retrieval, execution and link analysis algorithms to improve feature location in software," *Empir. Softw. Eng.*, vol. 18, no. 2, pp. 277–309, 2013, doi: 10.1007/s10664-011-9194-4.

[35] A. Prasanth, "Intelligent Recommendation System using Semantic information for Web Information Retrieval," Advances in Computational Sciences and Technology, vol. 10, no. 8, pp. 2367–2380, 2017, [Online]. Available: http://www.ripublication.com

[36] S. Irfan and S. Ghosh, "Web Mining for Information Retrieval," *Int. J. Eng. Sci. Comput.*, vol. 8, no. 4, pp. 17277–17283, 2018, [Online]. Available: http://ijesc.org/

[37] "Information Retrieval from Web Documents using Pattern Matching Algorithms," no. 12, 2019.

[38] Y. Djenouri, A. Belhadi, D. Djenouri, and J. C. W. Lin, "Cluster-based information retrieval using pattern mining," *Appl. Intell.*, vol. 51, no. 4, pp. 1888–1903, 2021, doi: 10.1007/s10489-020-01922-x.