International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Finding PageRank (PR) Using Stochastic Matrix and Multivariate Random Variable

Sreekanth Kavuri[1], Vedavathi Katneni[2]

GITAM (Deemed to be University)
Email: *Kavurikanth[at]gmail.com*
Email: *k.vedavathi[at]gmail.com*

Abstract: ***In the modern world, information is an incredibly important aspect that must not be neglected. In the field of information, several different procedures or algorithms are developed and put to use to extract the required data. Google uses an algorithm called PageRank (PR) to determine where websites should be placed in search engine results. One of the creators of Google, Larry Page, was respected for the naming of the company's ranking algorithm. The relevance of every website on the web is quantified using an algorithm called Page Rank. Many issues in web usage mining have yet to be overcome. The greatest way to quantify anything is via a rating. Because there is an ever-increasing wealth of high-quality information available online. The search engine's effectiveness in yielding useful results has a direct bearing on users' levels of happiness. It becomes more challenging without any help sorting through it all. The issue is that it is not simple to locate the desired information. This study proposes an algorithm for determining the rank of mined web pages that uses a transition matrix and random vector to address these problems. The analysis of the experiment reveals that there were 34 separate iterations of 9 pages, each with a unique PR.***

Keywords: Data Mining, Iteration, Random Vectors, Page rank, Markov Chain

1. Introduction

In today's world, information is an extremely crucial factor. In the information sector, several methods or algorithms are created and used to extract the needed data. Data mining (DM) has evolved into a natural technique that is used to extract information. The author defines DM as the process of concluding huge datasets [1]. The information that was mined for turned out to be beneficial. As a consequence, DM is often referred to as the extraction of awareness from data. Iteration is a key component of the DM process. DM in its more traditional form extracts information from prearranged formats such as tables, spreadsheets, or files. Now, as a consequence of the exponential expansion of the web and its contents, online text mining is in high demand. The amount of information available on the internet is already quite high and is only expected to keep rising. The material is covered in a variety of different ways and a wide variety of contexts. They have provided information on text mining, data mining, and web mining [2].

PageRank (PR) is the primary criteria that are used by the Google search engine to determine the order in which sites are shown in search engine rank [3] [4]. Figure 1 shows the web search engine's structure.

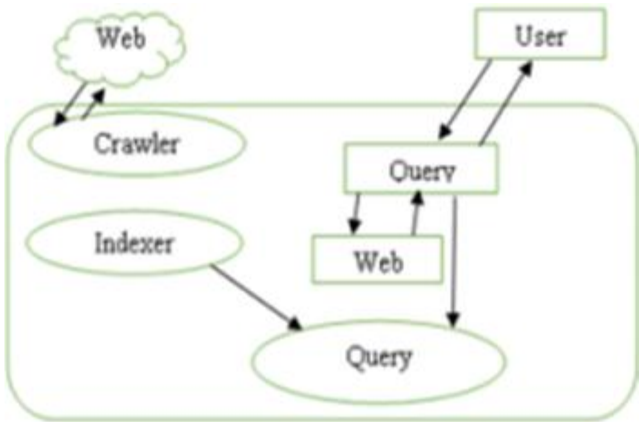


Figure 1: Search Engine Structure [5].

PR is an algorithm that assigns an importance score to each webpage on the web. To properly organize search results, an index like this one is used [6][7]. PR is a popular subject among Search Engine Optimization (SEO) professionals. The basis of PR is a mathematical formula that seems to be quite hard, but when simplified, it is fairly easy to grasp. Solutions, storage problems, and uniqueness are also estimated using the PR model [8]. Web developers need to be familiar with PR to improve a site's search engine rankings. There is a significant number of website offerings now accessible on the market. If a person had to sit down and calculate the PR for each website, this could be a very time-consuming task.PR could have a greater impact on the Internet if it is automated. This is because automation is making people's lives simpler. Google is primarily driven by an algorithm called PR. Larry Page and Sergey Brin came up with the idea for it when they were both graduate students at Stanford University. After that, in 1998, it was registered as a trademark for the company Google. PR's influence extends beyond the computer science field and into the commercial world. Every company's business aim is to be ranked higher in web page display, which is referred to as an SEO plan. The Page ranking is greatly influenced by the hyperlink structure as well as the SEO strategy.

Every algorithm has a mathematical foundation. It's fair to say that matrices and vectors are mostly responsible for many successes. The need for algorithms to rank websites has arisen due to the exponential growth in the quantity of available digital real estate. Google's page rank is unique in that it does not permit spam, which is websites that are programmed to manipulate search engine rankings in a manner that is against Google's criteria. Additionally, it emphasizes the page's significance when it is pointed to by other significant nodes. Linear equations and graphs play an integral role in the PR algorithm's operation [9].

In this study, the authors would discuss the mathematics behind the page ranking algorithm. Moreover, it details the present situation's issues and offers solutions.

1.1 Web Page Ranking Algorithm

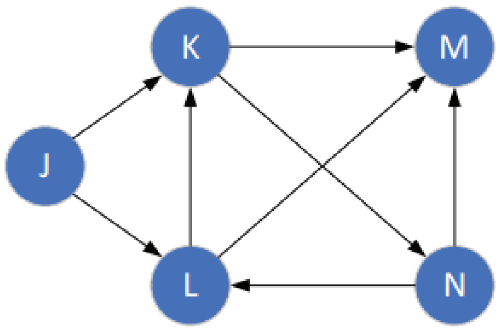
Rapid advances in network technology have resulted in a deluge of massive information resources throughout the whole Internet. The web search engine is quickly becoming the most popular method of accessing information. The basic purpose of search engines is to present users with relevant information. Therefore, several different web Page Ranking Algorithms are utilized to order the results of online page queries efficiently and effectively[10].

The need for best-quality results is the key cause in the invention of various web page ranking algorithms, Hyperlink Induced Topic Search (HITS), and Weighted PageRank (WPR) are distinct instances of web page ranking utilized in different situation. It is crucial to study the page rank algorithm utilized by Google because of the search engine's current prominence and the ripple effect it has on a large number of internet users.

1.1.1 HITS algorithm

The HITS algorithm can distinguish between two distinct types of websites that is hubs and authority. Authority sites are those that provide crucial information. Pages that function as resource listings and direct readers to relevant authorities are known as hubs. Therefore, a good hub page for a topic would link to many authoritative sites on that material, and a good authority page would be linked to many excellent hub pages on the same subject [11].

Figure 2 depicts the relationships between hubs and authorities. In this context, a page can serve as both a central information resource and a reliable authority. The cyclical link results in the development of an iterative algorithm known as HITS. The HITS algorithm ranks websites based on inbound and outbound connections. If a web page is linked to by numerous hyperlinks, it is known as an authority, and if the page points to several hyperlinks, it is known as a hub. The HITS algorithm is link-based. In the HITS algorithm, the ranking of a web page is determined by analyzing the textual contents of the website concerning a certain query. After gathering the web pages, the HITS algorithm focuses just on the web's structural elements while ignoring their textual contents [12-13].The following method is used to determine the hub and authority in this HITS implementation[14-15].



a(K)= h(J)+h(L)   a(L)=h(J)+h(N)
h(K)=a(M) +a(N)   h(L)=a(K)+a(M)

Figure 2: A simple example of calculating hub and authority values [13]

Let $a(p)$ and $h(p)$ denote the authority and hub scores of pages$p$. The sets of the referrer and reference pages of page p are denoted by $B(q)$ and $F(p)$, respectively. The HITS algorithm is separated into many steps:

(1) Compute $a(p)$ and $h(p)$ in a mutually reinforcing way as follows:

$$a(p) = \sum_{q \epsilon B(q)} h(q) \quad (1)$$
$$h(p) = \sum_{q \epsilon F(p)} a(q) \quad (2)$$

(2) Normalize the authority of all websites by dividing it by the page with the greatest authority:

$$a(p) = \frac{a(p)}{\max(a(p))} \quad (3)$$

To normalize the hub of all websites, divide it by the highest hub:

$$h(p) = \frac{h(p)}{\max(h(p))} \quad (4)$$

(3) Step 2 should be repeated until the difference in weight between the previous iteration and the current iteration is less than the chosen threshold, at which point the system has reached a stable state and $a(u)$ and $h(v)$ convergence.

1.1.2 Weighted Page Rank (WPR) Algorithm

Winpu Xing and Ali Ghorbani introduced the WPR method, a variation to the traditional Page Rank algorithm, in 2004. The issue of rank sink in the Page rank algorithm is solved by the WPR algorithm. Figure 3 indicates the illustration of WPR.

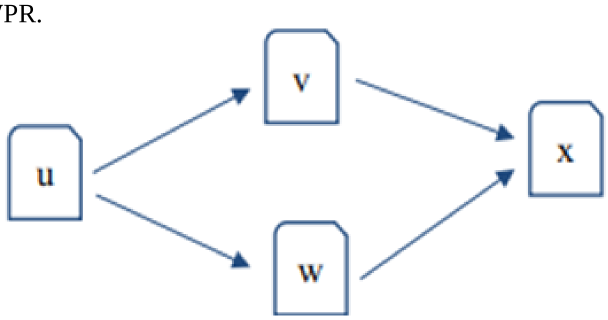


Figure 3: Illustration of in and out links [41]

According to the third equation, WPR prioritizes ranking higher-importance sites with higher rank scores rather than spreading the rank score evenly across all outbound links [16].

$$PageValue\ \alpha\ Popularity \tag{5}$$

$Win\,(v,u)$ and $Wout\,(v,u)$ represent the number of inward and outbound links, respectively, and are used to determine a page's credibility.As in equation (6), the value of web link (v,u) is the sum of the weights of all the inbound connections to website u and all the inbound links to all the reference websites for website v[17].

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p\epsilon R\,(v)} I_p} \tag{6}$$

The values I_u and I_p stand for the number of internal links to page u and page p, respectively. Wherever $R(v)$ provides the reference web page list of web page v and $Wout\,(v,u)$ (as in equation (7)) represents the weight of $link(v,u)$ computed based on the count of out-links of page u and the count of out-links of all reference web pages of page v[18].

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p\epsilon R\,(v)} O_p} \tag{7}$$

Where O_u denotes the pages that link out from page u, O_p denotes the pages that link out from page p, and $R(v)$ is the list of sites that link back to page v[19].

Modified WPR is:

$$PR(u) = (1-d) + d\sum_{v\epsilon B\,(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \tag{8}$$

Merits of Weighted Page Rank

- In comparison to the Page rank algorithm, the web pages return high-quality websites.
- It is more efficient than Page rank since the rank value of a page is distributed across its out-link sites based on the relevance of that page, regardless of whether it is included in a loop or not.

Demerits of Weighted Page Rank

It is more efficient than Page rank since the rank value of a page is distributed across its out-link sites based on the relevance of that page, regardless of whether it is included in a loop or not.

The following is the order in which the remaining portions of this work are presented: In part II, we look at the works that are linked to this topic. The background research was presented in part III, and a detailed issue formulation along with potential answers was presented in section IV. In section V, the architecture of the system and each of its components is detailed in detail. In the next part, section VI the emphasis is placed on the experimental results, coupled with a summary of the findings. At long last, authors get to the conclusion, as well as the future work

2. Literature of Review

Ma et al., (2022) [20] presented a straightforward and powerful solution called a Graph Mixed Random Network Based on PageRank (PMRGNN). The performance of the model can be enhanced by using a graph regularisation term that can glean more relevant information for classification results from neighbor nodes. Experiment findings on graph benchmark datasets reveal that the technique presented in this research beats various previously published Graph Neural Network (GNN) baselines on semi-supervised node categorization. PMRGNN consistently shows higher performance in studies examining over-smoothing and generalization. Classification visualization using this approach is more intuitive than previous ways.

Roth et al., (2022) [21] present the Personalized PageRank Graph Neural Network (PPRGNN), which expands the graph convolutional network to an infinite-depth model with the possibility of returning the neighbor aggregate to its original state in each iteration. Time complexity stays linear and memory complexity remains constant regardless of network depth, allowing it to scale effectively to vast networks. We demonstrate the efficacy of the method on several node and graph categorization tasks experimentally. In practically every scenario, PPRGNN performs better than competing approaches.

Shen et al., (2021) [22] provide a different preconditioning strategy to solve the PR model. This method converts the original PR eigen problem into a new one that is easier to solve. The authors describe a preconditioned version of the revised Arnoldi approach for solving this problem. Theoretically, the authors demonstrate that the preconditioned Arnoldi approach outperforms the refined Arnoldi technique in terms of execution efficiency and parallelism. This preconditioned approach has been shown to have a much quicker convergence speed than its usual counterpart in several numerical studies, particularly for challenging instances with high damping factors.

Lu et al., (2021) [23] demonstrate a method for estimating PR using a heat kernel on a small subgraph. In addition, the authors demonstrate that the proposed technique requires a sublinear number of computations in terms of the estimated size of the local cluster of interest and that it gives a decent approximation of the heat kernel PR with approximation errors capped by a probabilistic guarantee. The approximation heat kernel PR is used in a local clustering method, and numerical studies show that it reaches state-of-the-art performance.

Fushim et al., (2020)[24] developed a technique for efficiently calculating PR-based node ranking by first computing the precisely predicted transition matrix over all conceivable worlds and then running PR just once (p-avg approach). This differs from calculating scores for each graph separately and then averaging them to get a node's ranking (s-avg approach). We demonstrate that the suggested technique (p-avg approach) provides extremely high accuracy to the s-avg approach and could be a suitable replacement for it for highly ranking nodes.

Biganda et al., (2020) [25] provided an analysis of the three distinct forms of PR, including the original PR (created by Brin and Page), Lazy PR, and Random Walk with Backstep PR. Authors analyze the variants in terms of rank score convergence and consistency for various network architectures using PR's parameters, c (damping factor) and (backstep parameter). It also demonstrates additional

Paper ID: SR23513160827 DOI: 10.21275/SR23513160827

proportionality connections that allow us to derive the standard PR algorithm from the other two variations.

Zhang et al., (2019)[26] provided a technique for ranking the critical nodes of the power system that combines the effective resistance matrix with the PR algorithm. The suggested technique is enhanced by including the flow direction, which brings it closer to the real power system characteristics. Finally, the IEEE-39 node example has been verified in simulation to demonstrate that the enhanced critical node identification model has a specific recognition function for the nodes in the central position of the system topology, which can serve as a guide for the secure and stable operation of a real power grid.

He et al., (2018) [27] improved the accuracy and comprehensiveness of mining prominent users by developing the PR model and adding the evaluation of interaction and personal impact. The algorithm's logic and efficacy were tested through simulation using actual user data from a Sina microblog, grabbed using the programming language Python. The testing findings demonstrate the clear benefits of the approach in terms of accuracy and recall rate while looking for possible influential users. Table 1 indicates the comparison table of the literature of review.

Table 1: Comparison table of literature of review

Authors	Techniques Used	Outcomes
Maet al., (2022) [20]	PMRGNN	On three data sets (Cora, Cite Seer, PubMed), PMRGNN outperforms other top algorithms by 0.3%, 2.4%, and 0.2%, respectively.
Rothet al., (2022) [21]	PPRGNN	PPRGNN surpasses all other techniques by at least 1% on four of five datasets (MUTAG, PTC, COX2, PROTEINS, NCI1) and is the second-best performing model with competitive accuracy (85.5%) on the fifth.
Shenet al., (2021) [22]	Preconditioned Arnoldi and Refined Arnoldi	Numerical results show that this approach is much quicker than the refined Arnoldi for solving the PR issue over a large range of parameters. Additionally, the weighted-Arnoldi and the extrapolated-Arnoldi approaches can benefit from this preconditioning strategy.
Luet al., (2021) [23]	local clustering	Numerical studies show that the local clustering approach based on the approximate heat kernel PR delivers cutting-edge performance.
Fushim et al., (2020) [24]	PR	The suggested technique is S times quicker than the s-avg strategy and effectively is 10,000 times faster, and it only takes a single run of PR.
Bigandaet al., (2020) [25]	PR, lazy PR, random walk with backstep PR	Numerical measurements show that when c is small, rank scores for the top 10 vertices are the same using conventional and lazy Page Ranks. When c = 0.85 and is tiny (about 0.01), however, random walk with backstep performs comparably to standard PR.
Zhanget al., (2019) [26]	Effective resistance matrix and PR algorithm	The simulation results of the IEEE 39-Bus system show that the node ranking model utilizing the effective resistance matrix and PR algorithm can successfully identify the nodes with critical topological locations, particularly the terminal nodes, and so assist the avoidance of power system accidents.
Heet al., (2018) [27]	PR	The suggested approach takes around 10% longer (in the worst-case scenario), but as the number of users rises, the difference in running time between the new and old techniques remains the same.

3. Background Study

The internet is expanding at an extraordinary speed nowadays. It's worth noting that internet use has been expanding rapidly in recent years. The information available on the web in the form of pages is constantly evolving as new material is added and old pages are deleted. The web's information has become incredibly significant, and a vast quantity of information is hidden from the inside. Unfortunately, it has become very difficult to get the necessary data. Therefore, it is essential to conduct extensive online data mining concerning content, structure, and use. In general, authors can expect to get a list of relevant webpages for the requests from the search engines. A rating system is used so that users can easily advance up that list. There are a variety of ranking systems, but many of them are based on either content or links. In this research, the authors offer an algorithm to determine the order of the mined websites. The authors evaluate the suggested approach in light of two well-established mining methods: the page rank and HITS algorithms. In this study, they compare and contrast the various algorithms and provide a detailed analysis of their strengths and weaknesses. Moreover, the system prioritizes displaying the search engine-provided, highest-quality website at the top of the results [28].

4. Problem Formulation

Information is an extremely crucial factor in today's world. The information sector is always developing new methods and algorithms, which are then used to get the information that customers want. Information extraction via DM had evolved to become a natural technique. There is a significant amount of interest in analyzing data regarding web usage to gain a better understanding of web usage and to put that knowledge to use to provide better service to web users. There are currently several unresolved problems in the subject of web use mining. The best method for placing something on a scale is through a ranking. Given that the quantity and quality of online resources continue to grow exponentially. Customer satisfaction is based on the search engine's success in providing relevant results. Without any kind of aid in sifting through it all, it becomes progressively more difficult. The problem is that it's not easy to find what the user needs. It is the goal of this process to examine and assess the most well-liked and productive web pages available. The implementation of more stringent legislation has made the preservation of individuals' right to privacy the primary obstacle in a wide variety of practical applications. Web mining has developed into an effective method for mining information from the web. This interest is primarily driven by the continued growth of web-based applications, most notably electronic commerce. There has been an increase in the number of more stringent rules, which has made the protection of personal information a significant obstacle in many practical applications. It was for that reason that web mining had become such a potent tool for data collection on the World Wide Web. The approach used in this study effectively mines the web to get information in a way that is pleasant to users. The author then compares the proposed model to the standard page ranking approach, PR. However, the suggested model is valid for undirected networks lacking self-nodes, dangling edges, and null vertices.

5. Research Objectives

- To design a model for webpages ranking for a related query created by a user.
- To enhance the efficiency of the page ranking algorithm by implementing a transition matrix and random vector.
- To examine how the results of these transition matrix and random vector techniques differ when applied to PR.
- To quickly calculate the PR rating for a graph of uncertainty where every connection could be unknown.

6. Research Methodology

The idea of the designed architecture is analyzed and discussed in the framework of the research methodology. A technique that writers use to define how they intend to carry out their investigations is referred to as the "research methodology," and the word "research methodology" is referred to by the term "research methodology." It is a procedure that is rational and methodical, and it is intended to address a research issue. To ensure that their study yields reliable and valid data and that their aims and objectives are satisfied, authors often include an overview of their methodology. It considers the way the data is being gathered and evaluated, as well as the data itself, as well as where it could come from, and how it will be acquired. A query is generated by the user side. The search engine extracted the keywords from the query while considering the browsing history it shows some related webpages and hyperlinks further these webpages are ranked by constructing a graph then a transition matrix and a link matrix. A random vector is used for surfing the node. To protect the surfing from the dead end and spider traps a teleport matrix is also defined. Finally, based on calculated results, the ranking of the pages is completed.

6.1 Techniques Used

These technologies are used in the proposed methodology such as transition matrix, random vector, and page rank approach. It is possible to explain the process of transitioning between two states by making use of something called a transition matrix. A random vector is a vector of random variables X. PR is a method for determining the relative relevance of individual websites. It is utilized in situations in which events will either be more or less likely dependent on the events that have occurred in the past. These technologies are described below:

6.1.1 Transition Matrix

A Transition Matrix, also, known as a stochastic or probability matrix is a square (n x n) matrix representing the transition probabilities of a stochastic system (e.g. a Markov Chain). The size n of the matrix is linked to the cardinality of the State Space that describes the system being modelled[29].

The following formulas can be used to create a transition matrix (M):

$$M_{ij} = \frac{A_{ij}}{\sum j} \tag{9}$$

Where,
A = Link matrix
M_{ij} = Transition matrix [30].

The transition matrix stores all the data that pertains to the changes that occur while moving from one state to another. Figure 4 presents a diagrammatic representation of a generic transition matrix.

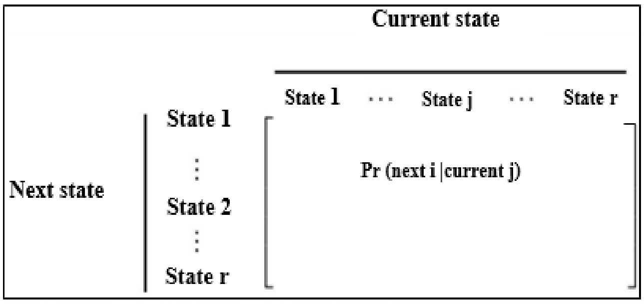


Figure 4: Generic Transition Matrix [31]

The adjacency matrix G could be used to describe the Web link structure. It looks like this: $G = g_{ij}$. The value of g_{ij} is 1 if and only if there is a connection from page j to page I, and 0 otherwise. This is how a transition matrix P is constructed from an adjacency matrix G, with entries consisting of:

$$p_{ij} = \begin{cases} \frac{1}{\sum_{k=1}^{n} g_{kj}}, & if\ g_{ij} = 1 \\ 0, & Otherwise. \end{cases} \tag{10}$$

According to the above equation, the probability of transitioning from one page to another is evenly dispersed among the outlines of that page. Keep in mind that the structure of the transition matrix P is the same as the structure of the adjacency matrix G. One possible mathematical formulation of the PR problem at this stage is the following linear system:

$$Ax = v, with\ A = (I - \alpha P) \tag{11}$$

where $0 < \alpha < 1$ is the damping parameter, which controls the weight allocated in the problem to the Web link structure, v is a probability distribution vector termed the customization or teleportation vector, and x is the unknown PR vector utilized for the ranking[32].

6.1.2 Random vector

A random vector is one whose magnitude is determined by the results of the investigation. Let a sample space beΩ. The set of K-dimensional real vectors R^K is a function from the sample space Ω to the random vector X [33].

$$X: \Omega \rightarrow R^K \qquad (12)$$

When working with several random variables at once, the usage of vector and matrix notations might prove to be quite helpful. This not only makes the formulae more concise but also enables us to utilize the linear algebra truths. Figure 5 An Icon of a Randomly Selected Vector.

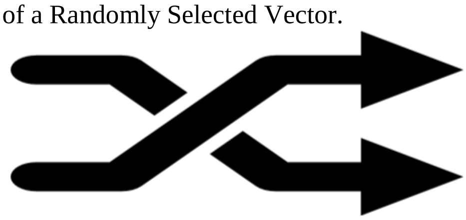


Figure 5: Icon of Random Vector [34]

This section of the study takes a cursory look at this possibility. The reader should have previous experience with matrix algebra before continuing with this subject. When there are n random variables, such as, $(X_1, X_2, X_3, \ldots X_n)$ it could organize them into a vector, denoted by the letter X:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ X_n \end{bmatrix} \qquad (13)$$

A random vector is denoted by the letter X. The variable X has a dimension count of n because it is made up of n different random variables.

6.1.3 Page Rank Algorithm

The PR algorithm, which is utilized by Google, is one of the most significant algorithms. PR's basic tenet is that a website's relative relevance can be inferred from the prominence with which other websites connect to it. If theyproduce a web page i that has a hyperlink attached to page, j, then page j is deemed as important. On the other hand, page j receives a backlink from page k (such as www.google.com), authors can argue that k transfers its control to j. (i.e., k asserts that j is important). They can give a rank to each page iteratively according to the number of pages that link to it[35].

A PR Algorithm contemplates more than 25 billion web pages on the www to allocate a rank score [36]. A simplified version [37] of Page Rank is described in Equation 14[38]:

$$PR(u) = C \sum_{V \epsilon B(u)} PR(v)/N_v \quad (14)$$

Here, u stands for a webpage, $B(u)$ is the collection of pages that link to u, $PR(u)$, and $PR(v)$ are rank scores for site u and page v, respectively, N_v is the number of outgoing pages v that link elsewhere, and C is a normalization factor. Figure 6 indicates the Distribution of PR.

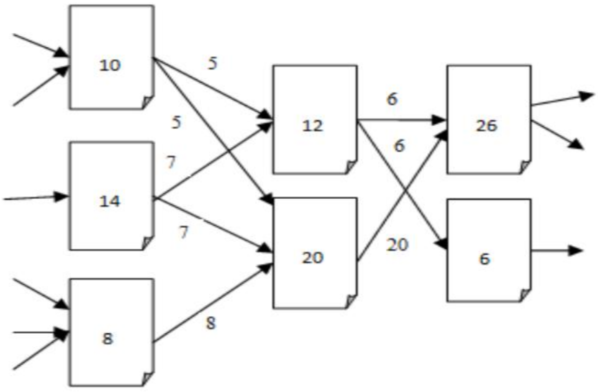


Figure 6: Distribution of PR [39]

The algorithm was updated once it was realized that not all Internet users like to click on direct linkson WWW. The modified version is given in Equation 15 [40]:

$$PR(u) = (1 - d) + d \sum_{V \epsilon B(u)} PR(v)/N_v \quad (15)$$

Here $'d'$ is a damping factor, often set to 0.85, and it could be thought of as the probability of users' following the links, with $(1 - d)$ being the page rank distribution from non-directly connected pages [10].

Merits of PR

- As it is a query-independent method that precomputes the rank score, it requires extremely little time.
- This approach is more feasible because the rank score is calculated during indexing, not during query processing.
- It returns significant pages because rank is based on a page's popularity.
- It is more resistant to link spam since the full web graph is used to determine a page's rank value instead of just a tiny sample[41].

Demerits of PR

- It gives more value to older pages, as a new page, even a good one, won't have many connections unless it's part of an established website or a loop of web pages.
- The content of websites is not considered, hence the relevance of the returned results to the user query is low at best.
- Presence of Dangling connections. This happens when a page has a link such that the hypertext refers to a page with no outbound connections.

7. Proposed Methodology

Search engines utilize several Algorithms to evaluate websites and determine which one is most relevant to a user's search query, with the ranking representing a significant place on a scale. The user submits a query to the search engine to obtain some information from the website, and the search engine then returns some results in response to the user's question. As a consequence of this, the author ends up with several web pages that are connected by hyperlinks. It is necessary to provide a ranking to these final web pages. Therefore, to accomplish this goal, the web pages and the links are regarded as a graph and ranked according to the methodology that is detailed below. Figure 7 shows the flowchart of the proposed methodology.

The proposed methodology is completed in 6 steps. These steps are given below:

Step 1: Query generation
A query generation, which is also known as a data preview, is a copy of the data in the persistent store. A query is requested from the user at the beginning of the procedure to search for a certain issue. At the beginning of the process of searching for a particular problem, the user would be asked to provide a query.

Step 2: Extracted keywords
In unstructured text data, important phrases can be found using a technique called keyword extraction, which is used to automatically pluck out single keywords or groupings of two or more words. Step 2 involves the generation of a question, after which the search engine would extract similar terms based on the user's previous browsing history.

Step 3: Searching webpages

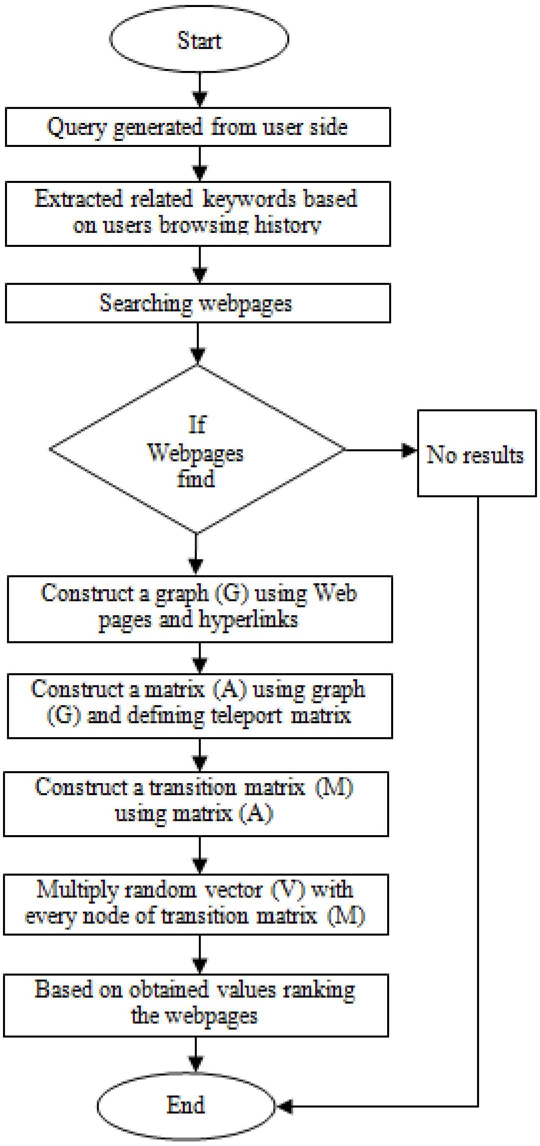


Figure 7: Flowchart of the proposed methodology.

The words or phrases a user enters with their massive databases of information to provide a list of results that are relevant to the user's query. After extracting the related keywords in this step search engine search the related webpages and hyperlink if webpages are not extracted then the search engine shows no result found and if the webpages are found then the process move to the next step.

Step 4: Construct a graph (G)
In step 4, once the webpages and hyperlinks linked to the keywords have been retrieved, a graph (G) is formed by utilizing the webpages as vertices (V), and the hyperlinks as edges (E).

Step 5: Construct a transition matrix (M)
After the construction of the graph (G) in this step, a transition matrix (M) is constructed. With the help of graph (G), a link matrix (A) is generated by making links to each vertex and after this, a transition matrix (M) is generated using the matrix (A).

Step 6: Role of a random vector
In step 6, after the construction of a transition matrix to calculate the rank of the webpages a random vector is taken for surfing. The random vector is multiplied with each node of the graph i.e.,

$$v_i = M.v \tag{7}$$

Where i vary from 1 to n, no. of nodes, M is the transition matrix and v is the random vector. After multiplying the random vector with all nodes, the rank of the pages is obtained.

Proposed Algorithm:
The proposed algorithm is to design K means Ad hoc On-Demand Multipath Distance Vector (K – AOMDV) Routing Protocol

Start

Step 1: Problem searches require a question. Persistent storage data previews are query-generation copies.
Step 2: **N** number of **Consecutive words** can extract **keywords** from **unstructured text**.
Step 3: The **search engine** would pull **comparable phrases** from the **user's browsing history**.
Step 4: **Predictive search** can be used to search **next word** using the **shortest path algorithm** and inverted indexing to show words suggestion

Step 5: The terms a user types with their **huge databases** of information to get **appropriate results**.
Step 6: **Search engines** search relevant **web pages** and **hyperlinks** after **extracting keywords**.
Step 7: If webpages are not extracted, the search engine returns no results, but if they are, the process continues.

Step 8: Take a graph **(G)** as **a keyword** with vertices **(V) as websites** and edges **(E) as hyperlinks** between them.
Step 9: Form a link matrix containing all linkages to each vertex (A).
Transition matrix (M):

$$M_{gv} = \frac{A_{gv}}{\Sigma_v}$$

Paper ID: SR23513160827 DOI: 10.21275/SR23513160827

g is the **Keyword**,
v is the **WebSite**, and
Σ all matched websites with ***Keywords***

Step 10: Surfing with a **random vector** ranks web page. The random vector is multiplied by each graph node.

$\boldsymbol{v_i = M.v}$

M is the **Transition Matrix**,
v is the **Random Vector**, and
i ranges from ***1*** – ***n***, the *number of nodes*.

Step 11: After multiplying the random vector with all nodes, the page rank is obtained.

End

8. Result and Implementation

The outcomes of the procedures that have been detailed up to this point are discussed in this section. The total number of web pages and the number of hyperlinks connecting those sites could be included in the findings. These web pages that have been produced need to be ranked. Therefore, to obtain better results, the web pages and the connections are regarded as a graph and ordered according to the following methodology:

1) First, consider a graph (G) where the vertices (V) are web pages, and the edges (E) are the hyperlinks (H) between them.
2) Create a matrix that contains the number of connections connecting each vertex and call it a link matrix (A).
3) Form a transition matrix (M), where

$$M_{ij} = \frac{A_{ij}}{\Sigma j}$$

4) Take a vector v and start a random surfer at a vertex.
5) Multiply v with M and get v1
6) Step 5 should be performed the same number of times as the number of nodes (minimum number of times).
7) If the vector contains duplicate values, we must return to Step 5 until we have a set of distinct numbers.
8) V will be among the most prestigious pages.

A large number of iterations are performed to get an approximation of the findings. Page and Brin advised 100 iterations for accurate approximation, given the web's countless billions of pages [42]. However, there are only 9 pages total in Figure 8, and only 34 iterations have been completed. The iteration results are shown in Table 2.

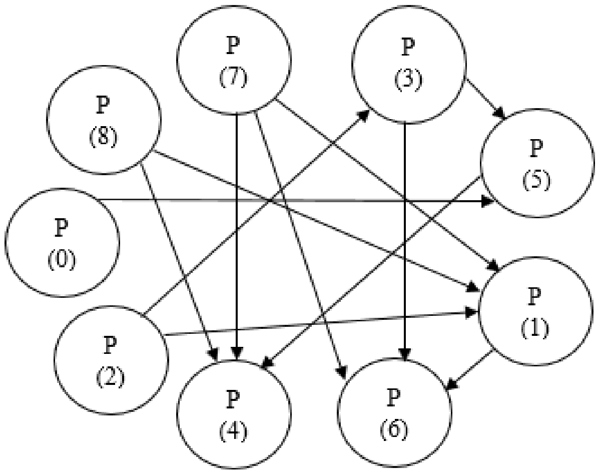


Figure 8: Example Graph G

Table 2: Results of PR Iterations

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
0	1	1	1	1	1	1	1	1	1
1	0.02469136	0.09876543	0.0617284	0.11728395	0.13580247	0.22839506	0.13580247	0.11728395	0.08024691
2	0.03840878	0.12277092	0.04663923	0.11454047	0.10631001	0.25034294	0.11865569	0.09602195	0.10631001
3	0.03848499	0.11621704	0.05128791	0.1013565	0.12764822	0.24573236	0.13039171	0.09724127	0.09163999
4	0.03810818	0.11535758	0.05093651	0.10721604	0.12022219	0.24846737	0.12346102	0.0942988	0.10193229
5	0.03808513	0.11497749	0.05078786	0.10470706	0.12321782	0.24751854	0.12497907	0.09753082	0.09819622
6	0.03833882	0.11490086	0.05103386	0.10539244	0.12209176	0.24756767	0.12426204	0.09646482	0.09994773
7	0.03822583	0.11477733	0.05100544	0.10516344	0.12262163	0.2474394	0.09705867	0.09705867	0.09927171
8	0.03827756	0.11481089	0.05101951	0.10525913	0.12236919	0.24747139	0.1243331	0.09687085	0.09958838
9	0.03826025	0.11479099	0.05101944	0.10521437	0.12248586	0.24744201	0.12437611	0.09696613	0.09944484
10	0.03826757	0.11479773	0.05102099	0.10523599	0.12243293	0.24745261	0.12435611	0.09692557	0.0995105
11	0.03826424	0.11479494	0.0510201	0.10522677	0.12245618	0.24744749	0.12436467	0.0969449	0.09948071
12	0.03826582	0.1147963	0.05102057	0.10523076	0.12244578	0.24744961	0.12436114	0.09693611	0.09949391
13	0.03826508	0.11479571	0.05102035	0.10522908	0.12245042	0.24744868	0.12436271	0.09694	0.09948797
14	0.03826541	0.114796	0.05102043	0.10522982	0.12244833	0.24744911	0.12436203	0.09693824	0.09949062
15	0.03826526	0.11479588	0.0510204	0.10522949	0.12244927	0.24744892	0.12436234	0.09693902	0.09948943
16	0.03826533	0.11479594	0.05102041	0.10522964	0.12244885	0.24744901	0.1243622	0.09693867	0.09948996
17	0.0382653	0.11479591	0.05102041	0.10522957	0.12244904	0.24744897	0.12436226	0.09693882	0.09948972
18	0.	0.16666667	0.	0.16666667	0.	0.41666667	0.25	0.	0.
19	0.	0	0.13888889	0.08333333	0	0.40277778	0.31944444	0.	0.
20	0	0.13425926	0.02777778	0.10648148	0	0.40972222	0.27546296	0	0
21	0	0.12731481	0.05401235	0.09182099	0	0.40933642	0.28510802	0	0.02314815
22	0	0.125643	0.05015432	0.09503601	0	0.41210134	0.28234311	0.00771605	0.02700617
23	0	0.12579304	0.05088306	0.09411437	0	0.41166195	0.28346836	0.00900206	0.02507716
24	0.	0.12586091	0.05029007	0.09448945	0	0.41192451	0.28363447	0.00835905	0.02544153
25	0	0.12604131	0.05043415	0.09454482	0	0.41176227	0.2835919	0.00848051	0.02514503
26	0	0.12604558	0.05039545	0.09453063	0	0.41177949	0.2836501	0.00838168	0.02521707
27	0.	0.12606025	0.05042088	0.09455003	0	0.41175944	0.28360599	0.00840569	0.02519772

28	0.	0.12605201	0.05041932	0.09453533	0	0.41176433	0.28361932	0.00839924	0.02521044
29	0	0.12605155	0.05042082	0.09453977	0	0.41176382	0.2836109	0.00840348	0.02520966
30	0	0.12605022	0.0504204	0.09453697	0	0.41176469	0.28361409	0.00840322	0.02521041
31	0	0.11535758	0.05093651	0.10721604	0	0.24846737	0.12346102	0.0942988	0.10193229
32	0	0.12605035	0.05042018	0.09453767	0	0.41176473	0.28361355	0.0084034	0.02521011
33	0	0.12605041	0.05042015	0.09453785	0	0.41176472	0.28361341	0.00840337	0.02521009
34	0.	0.12605041	0.05042015	0.09453785	0	0.41176472	0.28361341	0.00840337	0.02521009

Result 1: Iteration 1

An estimate of the PR might emerge for the first iteration. P (0), P (1), P (2), P (3), P (4), P (5), P (6), P (7), and P (8) are the nine different web pages that are present in the first iteration. Page 5 has the highest page rank, next by Page 6, and page 4 both have equal value and then Page 3, Page 7, Page 1, Page 8, Page 2, and Page 0 in that sequence as shown in below table 3.

Table 3: Initial PR Iteration Outcomes

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
1	0.02469136	0.09876543	0.0617284	0.11728395	0.13580247	0.22839506	0.13580247	0.11728395	0.08024691

Consider graph G (1) (shown in Figure 9) with nine vertices and sixteen edges. This would mean that the search engine we used returned a total of nine web pages as a result of the query, and we could further imagine that these nine web pages are connected by a total of sixteen links:

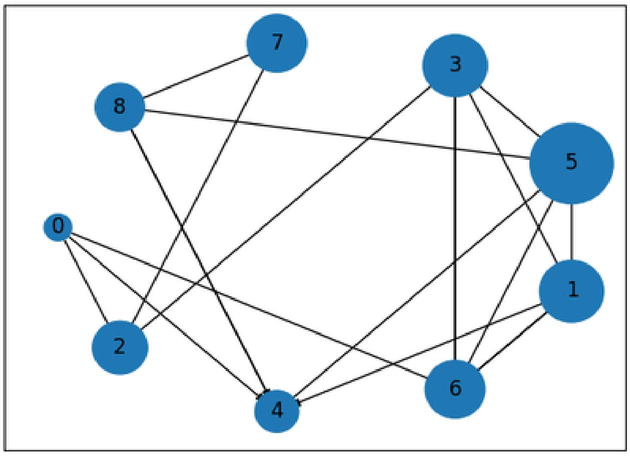


Figure 9: Graph G (1) of Iteration 1.

Result 2: Iteration 18

After the first iteration, every graph of the subsequent iterations is the same, however, there is a change in graph 18. For just the 18th iteration, an estimate of the page rankings could become accessible. The eighteenth iteration has nine distinct webpages, labelled P (0), P (1), P (2), P (3), P (4), P (5), P (6), P (7), and P (8). Page 5 is ranked highest, next the Pages 3 and 1, which are tied for second place, Pages 6 and 7, which are tied for third place, and finally Pages 2, 7, 8, 4, and 0 which are all equal in value. Table 4 displays the sequential values of all the pages.

Table 4: The eighteenth Iteration of PR's Results

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
18	0	0.16666667	0	0.16666667	0	0.41666667	0.25	0	0

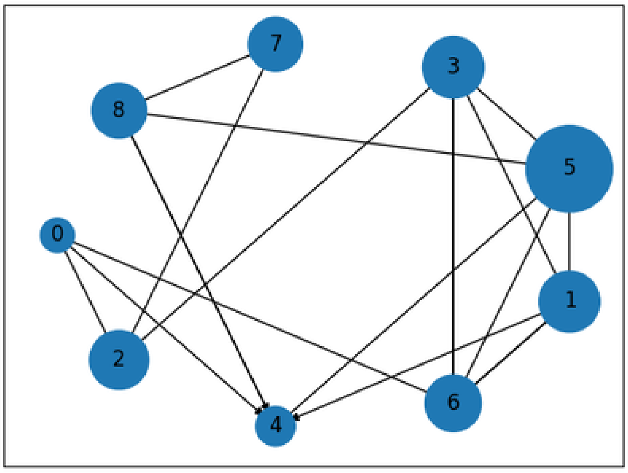


Figure 10: Graph G (18) of Iteration 18

Figure 10, graph G (18) has nine vertices and sixteen edges. In this case, the search engine would have returned a total of nine web pages in answer to the query and can even pretend that the sixteen links inside these nine websites are legitimate.

Result 3: Iteration 20

After the eighteenth iteration, the graph of the nineteenth iteration remains the same, but there is a difference in graph 20. It's possible that a rough estimate of the page rankings could emerge in the twentieth cycle. P (0), P (1), P (2), P (3), P (4), P (5), P (6), P (7), and P (8) are the titles of the first nine pages of the website in the twentieth iteration. Table 5 shows that Page 5 is the most frequent, followed by Page 6 and then Pages 1, 3, 2, and finally Pages 7, 8, 4, and 0 which have the same rank.

Table 5: Twentieth Cycle of PR iteration

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
20	0	0.13425926	0.0277778	0.10648148	0	0.40972222	0.27546296	0	0

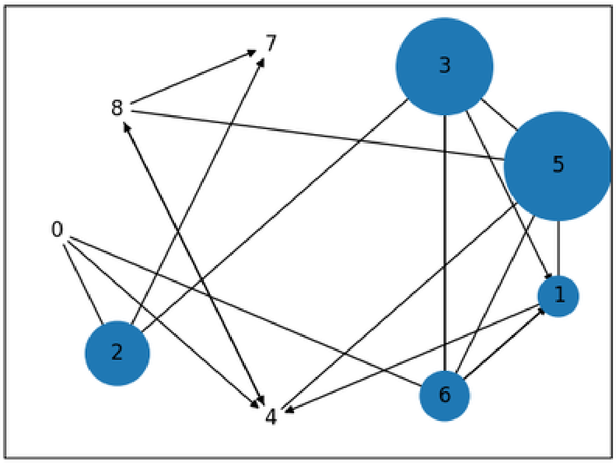


Figure 11: Graph G (20) of Iteration 20

Consider graph G (20) (shown in Figure 11), which has nine vertices and sixteen edges in total. This would imply that the search engine that was used returned a total of nine web pages because of the query, and it could go one step further and pretend that these nine web pages are related to one another by a total of sixteen links.

Result 4: Iteration 21

For the twenty-first cycle, a rough approximation of the page rankings should emerge. The twenty-first iteration has nine distinct webpages, labeled P (0), P (1), P (2), P (3), P (4), P (5), P (6), P (7), and P (8). The top-ranked page is page 5, next then page 6, after which page 1, then page 3, page 2, page 8 respectively, and finally page 4,page 7, and page 0 which havethe same rank. In table 6 below, you'll see a sequence of values for each page.

Table 6: The twenty-first Iteration of PR's Results

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
21	0	0.12731481	0.05401235	0.09182099	0	0.40933642	0.28510802	0	0.02314815

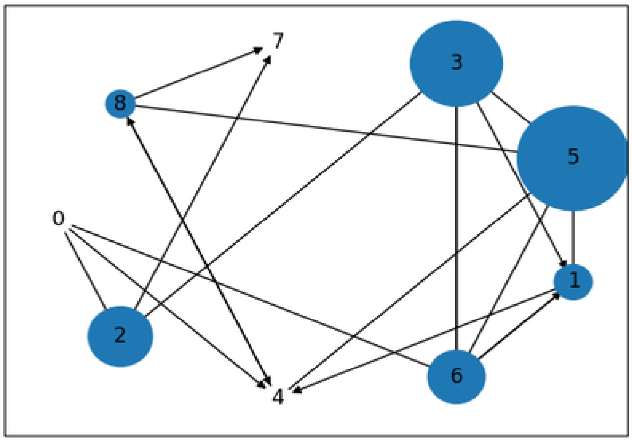


Figure 12: Graph G (21) of Iteration 21

The graph represented by the number G (21) in Figure 12 has nine nodes and sixteen edges. This would indicate that the search engine produced a total of nine online sites in answer to the query, and it could go as far as to assume that the sixteen links that exist among these nine web pages are real.

Result 5: Iteration 34

During the twenty-first iteration, every graph of the succeeding iterations looks the same, whereas the thirty-fourth is the last iteration. In the thirty-fourth cycle, perhaps some approximation of the PR could emerge. In the initial version, there are nine distinct web pages labeled P (0), P (1), P (2), P (3), P (4), P (5), P (6), P (7), and P (8). As seen in table 7, Page 5 has the greatest page rank, next by Page 6 and then Pages 1, 3, 2, 8, and 7 have the greatest rank, after those remaining pages 4, and 0 have the same rank.

Table 7: Thirty-fourth Cycle of PR iteration

Iteration	PR (0)	PR (1)	PR (2)	PR (3)	PR (4)	PR (5)	PR (6)	PR (7)	PR (8)
34	0	0.12605041	0.05042015	0.09453785	0	0.41176472	0.28361341	0.00840337	0.02521009

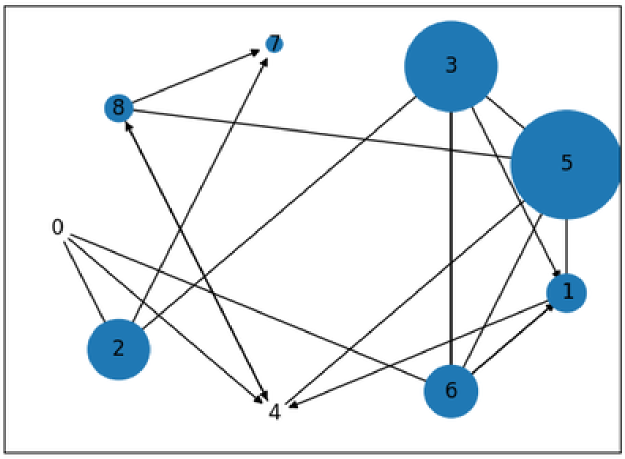


Figure 13: Graph G (34) of Iteration 34

Consider graph G (34) (shown in Figure 13), which has nine vertices and sixteen edges in total. This would imply that the search engine that was used returned a total of nine web pages because of the query, and it could go one step further and pretend that these nine web pages are related to one another by a total of sixteen links.

After 34 iterations the correct approximation of PR will emerge, where page 5 has the highest page rank, then page 6 will come, and page 1, page 3, page 4, page 8, and page 7 consecutively.

9. Conclusion and Future Scope

There is a growing interest in analyzing online user data to better understand web usage and apply the information to provide better to consumers, particularly as web-based apps, and notably electronic commerce, continue to expand in popularity. To determine a site's PR, webmasters start with the premise that every one of the billions of websites in existence on the Internet has some significance (0 being the least and 10 being the most important).A website's PR is determined by the quality and quantity of links leading to it. This paper's author uses a transition matrix and a random vector to calculate the PR.

The study of the results indicates that there have been 34 iterations of nine pages, each of which has a distinct rank. Page 5 has the highest page rank, followed by page 6, and then page 1, page 3, page 4, page 8, and page 7 in that order. Google no longer uses the moniker PR for its ranking algorithm, but the method is just as crucial as ever for how search engine results are organized. In the future, the method will be developed to accommodate the many edge circumstances mentioned.

References

[1] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.

[2] Shadab Irfan, Subhajit Ghosh, "Web Mining for Information Retrivel" in International Journal of Engineering Science and Computing, 2018.

[3] D.R.W. Holton, I. Nafea, M. Younas, I. Awan. A class-based scheme for Ecommerce web servers: Formal specification and performance evaluation. Journal of Network and Computer Applications, Volume 32, Issue 2, March 2009, Pages 455-460.

[4] Olivier Andrieu, Réussir son référencement web: Stratégie et techniques SEO, Eyrolles - Edition 2014 (16 décembre 2013)

[5] https://en.wikipedia.org/wiki/Search_engine_%28computing%29.

[6] Alexander Nazin, Boris Polyak, Adaptive randomiséealgorithme pour trouver le vecteur propre de la matricestochastique avec application de PageRank (48e Conférence IEEE-Décembre 16-18, 2009).

[7] Preethi, E., and R. Arumugam. "APPLICATIONS OF STOCHASTIC MODELS IN WEB PAGERANKING, 2017."

[8] M. Bianchini, M. Gori and F. Scarselli, "Inside PageRank", ACM transactions on Internet Technology, 2005.

[9] Dode, Albi, and Silvester Hasani. "PageRank algorithm." *OSR J. Comp. Eng. (IOSR-JCE)* 19, no. 01 (2017): 01-07.

[10] Singh, Charanjit, and S. K. Kautish. "Page Ranking Algorithms for Web Mining: A Review." *International Journal of Computer Applications* 975: 8887, 2015.

[11] https://onlineresource.ucsy.edu.mm/bitstream/handle/123456789/841/11086.pdf?sequence=1&isAllowed=y.

[12] Rekha Jain, Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer application,Vol 13, Jan 2011.

[13] The PageRank/HITS algorithm, Joni Pajarinen, March 19, 2008.

[14] Lei, Xiujuan, Siguo Wang, and Fangxiang Wu. "Identification of essential proteins based on improved HITS algorithm." *Genes* 10, no. 2 (2019): 177.

[15] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.

[16] Dilip Kumar Sharma and A.K Sharma, "A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, Vol.02, Issue 08, 2010.

[17] Kaushal K., Abhaya, Fungayi D., 2013: 'PageRank algorithm and its variations: A Survey report', IOSR Journal of Computer Engineering (IOSR-JCE), 14.

[18] Meng Cui, Songyun Hu, 2011: 'Search Engine Optimization Research for Website Promotion', IEEE computer society.

[19] Sharma, Prem Sagar, Divakar Yadav, and Pankaj Garg. "A systematic review on page ranking algorithms." *International Journal of Information Technology* 12, no. 2 (2020): 329-337.

[20] Ma, Qianli, Zheng Fan, Chenzhi Wang, and Hongye Tan. "Graph Mixed Random Network Based on PageRank." *Symmetry* 14, no. 8 (2022): 1678.

[21] Roth, Andreas, and Thomas Liebig. "Transforming PageRank into an Infinite-Depth Graph Neural Network." *arXiv preprint arXiv:2207.00684* (2022).

[22] Shen, Zhao-Li, Hao Yang, Bruno Carpentieri, Xian-Ming Gu, and Chun Wen. "A Preconditioned Variant of the Refined Arnoldi Method for Computing PageRank Eigenvectors." *Symmetry* 13, no. 8 (2021): 1327.

[23] Lu, Zhenqi, Johan Wahlström, and AryeNehorai. "Local clustering via approximate heat kernel PageRank with subgraph sampling." *Scientific Reports* 11, no. 1 (2021): 1-19.

[24] Fushimi, Takayasu, Kazumi Saito, Kouzou Ohara, Masahiro Kimura, and Hiroshi Motoda. "Efficient computing of PageRank scores on exact expected transition matrix of large uncertain graph." In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 916-923. IEEE, 2020.

[25] Biganda, Pitos Seleka, Benard Abola, Christopher Engström, John Magero Mango, Godwin Kakuba, and Sergei Silvestrov. "Exploring the relationship between ordinary PageRank, lazy PageRank and random walk with backstep PageRank for different graph structures." *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods* 5 (2020): 53-73.

[26] Zhang, Zhiyu, and Kounan Zhang. "Identification of critical nodes in power system based on graph theory and pagerank algorithm." In *2019 IEEE 8th International Conference on Advanced Power System Automation and Protection (APAP)*, pp. 213-217. IEEE, 2019.

[27] He, Fuyun, Binbin Huang, Yanting Shen, and Zebin Yang. "User Mining Algorithm Based On PageRank Modeling Improvement and Comprehensive Influence Evaluation." In *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pp. 398-402. IEEE, 2018.

[28] Kavuri Sreekanth, Katneni Vedavathi "Finding Page Rank using Transition Matrix and Random Vector." In

2019 International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. 23-27. IJITEE 2019.
[29] https://www.openriskmanual.org/wiki/Transition_Matrix.
[30] Nawaz, Falak, Mehdi RajabiAsadabadi, Naeem Khalid Janjua, Omar Khadeer Hussain, Elizabeth Chang, and MortezaSaberi. "An MCDM method for cloud service selection using a Markov chain and the best-worst method." Knowledge-Based Systems 159 (2018): 120-131.
[31] https://www.google.com/search?q=transition+matrix&rlz=1C1RXQR_enIN1021IN1021&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjtwMa0qZT8AhWrTWwGHVQ4D7QQ_AUoAXoECAEQAw&biw=1366&bih=657&dpr=1#imgrc=Nz0UYoKKKk2KmM.
[32] Shen, Zhao-Li, Ting-Zhu Huang, Bruno Carpentieri, Xian-Ming Gu, and Chun Wen. "An efficient elimination strategy for solving PageRank problems." Applied Mathematics and Computation 298 (2017): 111-122.
[33] https://www.statlect.com/fundamentals-of-probability/random-vectors.
[34] https://www.veryicon.com/icons/miscellaneous/practical-life-icon/random-3.html.
[35] Dutta, Er Manika, and Kishori Lal Bansal. "Improvement in Weighted Page Rank Algorithm using Efficiency and Precision, 2017."
[36] Kleinberg J., 1998. Authorative Sources in a Hyperlinked Environment". Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
[37] L. Page, S. Brin, R. Motwani, and T. Winograd, 1999. The Pagerank Citation Ranking: Bringing order to the Web. Technical report, Stanford Digital Libraries, SIDL-WP1999-0120.
[38] Srinath, K. R. "Page Ranking Algorithms–A Comparison." *International Research Journal of Engineering and Technology (IRJET)* (2017).
[39] Neelam Duhan, A.K Sharma and Komal Kumar Bhatia, “Page Rank Algorithms: A Survey”, IEEE International Advance Computing Conference (IACC), March 2009.
[40] Chowdhary, Aditi, and Arvind Kumar. "Study of web page ranking algorithms: a review." *Acta Inform. Malaysia* 3, no. 2 (2019): 01-04.
[41] Dutta, Er Manika, and Kishori Lal Bansal. "Improvement in Weighted Page Rank Algorithm using Efficiency and Precision."
[42] Nguyen Thon Da, Tan Hanh, Pham Hoang Duy, “Improving web page access predictions based on sequence prediction and PageRank algorithm”, in Interdisciplinary of Information, Knowledge, and Manegemnt”, January 2019.

Author Profile

Sreekanth Kavuri is presently a Research Scholar in the Computer Science Department, GITAM Institute of Science, GITAM (Deemed to be University), Visakhapatnam. Completed his Post Graduation from JNTU Hyderabad. Completed his pre-Ph.D. from GITAM, Visakhapatnam. Presently working as Assistant Professor in KL Deemed to be University. Previous an employee in AP state government university Taught many subjects in computer science and engineering department and Information Technology department.

Dr. Vedavathi Katneni is the Head of the Department, Computer Science, GIS, GITAM (Deemed to be University), Visakhapatnam. Had done her Ph.D. from Andhra University, Visakhapatnam. She is a life member of Indian Science Congress Association and Indian Society of Technical Education. Presently in GITAM (Deemed to be University), she was awarded with many awards, from women empowerment cell in GITAM (Deemed to be University) for her contribution towards women empowerment and initiatives taken towards empowering women.