DESIGN THINKING AND INNOVATION

ON

PAGE RANKING IMPROVEMENT ON WEB

Submitted by:

Harshita Bhardwaj,Harshita rout,navdeep
1/21/FET/BCS/135,22/FET/CS (L)/001,22/FET/CS (L)/005

Under the Guidance of

Bhanu Dwivedi
PROFESSOR

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
Computer Science & Engineering

Faculty of Engineering & Technology

MANAV RACHNA INTERNATIONAL INSTITUTE OF
RESEARCH AND STUDIES, Faridabad
NAAC ACCREDITED ‘A’ GRADE
June-july,2022

ABSTRACT

With the tremendous growth of information available to end users through the Web, search engines come to play ever a more critical role. Nevertheless, because of their general purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. Next generation Web architecture, represented by Semantic Web, provides the layered architecture possibly allowing to overcome this limitation. Several search engines have been proposed, which allow to increase information retrieval accuracy by exploiting a key content of Semantic Web resources, that is relations. However, in order to rank results, most of the existing solutions need to work on the whole annotated knowledge base. In this paper we propose a relation-based page rank algorithm to be used in conjunction with Semantic Web search engines that simply relies on information which could be extracted from user query and annotated resource. Relevance is measured as the probability that retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition. The PageRank algorithm assigns a score to each page based on the number and quality of links pointing to that page. The algorithm takes into account the link structure of the web and the PageRank scores of the pages linking to a particular page. The algorithm is iterative and converges to a stable value for each page. The PageRank algorithm has become a cornerstone of modern search engine optimization and is used by many search engines besides Google. The algorithm has also been criticized for being too easily manipulated. Nonetheless, the PageRank algorithm remains a powerful tool for ranking web pages and is likely to remain an important part of search engine technology for the foreseeable future

INTRODUCTION

The PageRank algorithm is based on the concept of citation analysis, which is used in academic research to measure the impact and importance of scholarly articles. The idea behind citation analysis is that articles that are cited more frequently by other articles are likely to be more important and influential.

Similarly, the PageRank algorithm considers links from other pages to be "votes" for the importance of a particular page. However, not all votes are created equal. Links from pages with higher PageRank scores are considered to be more valuable than links from pages with lower PageRank scores. Therefore, a page that has a small number of links from high-quality pages may be ranked higher than a page with a large number of links from low-quality pages.

The PageRank algorithm is an iterative algorithm, meaning that it is applied multiple times to refine the PageRank scores for each page. The algorithm takes into account the link structure of the web and the PageRank scores of the pages linking to a particular page. The algorithm converges to a stable value for each page, which is used to rank the pages in search engine results.

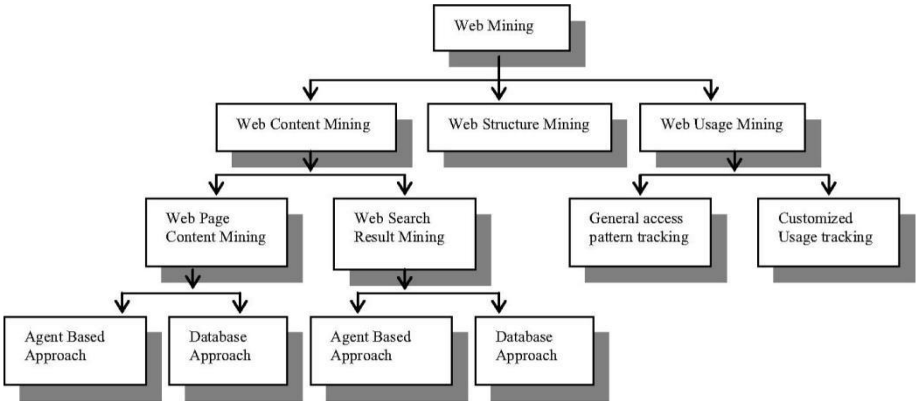
The PageRank algorithm has become a cornerstone of modern search engine optimization (SEO) and is used by many search engines besides Google. However, the algorithm has also been criticized for being too easily manipulated by webmasters who use tactics like link spamming to artificially boost their PageRank scores. Nonetheless, the PageRank algorithm remains a powerful tool for ranking web pages and is likely to remain an important part of search engine technology for the foreseeable future.

The PageRank algorithm has become an essential component of modern search engine optimization (SEO) and is used by many search engines, including Google. The algorithm has also undergone numerous iterations and improvements since its initial development, as search engines continually strive to provide more accurate and relevant search results.

Despite its usefulness, the PageRank algorithm has also been criticized for being too easily manipulated by webmasters using tactics like link spamming to artificially boost their PageRank scores. Nonetheless, the PageRank algorithm remains a crucial tool for ranking web pages and is likely to remain a fundamental part of search engine technology for the foreseeable future.

WEB MINING

Web structure mining and web usage mining. Web content mining involves the extraction of useful information from the web content. This includes extracting the text, images, audio, and video from the web pages. Data mining techniques such as clustering and classification are used to discover the hidden patterns and relations among the web content. Web structure mining is the process of extracting the structure of the web. This includes extracting the links between web pages and discovering the type of relationship between them. Web usage mining is the application of data mining techniques to discover the user's access patterns from the web server logs. It helps in understanding the user's navigation behaviour and personalizing the web services. Therefore, data mining and web mining techniques have a major role in improving the web services and discovering useful information from large databases.



DATA MINING

Data mining is the process of extracting intriguing information or patterns from sizable databases that are non-trivial, implicit, previously undiscovered, and possibly beneficial. Web mining is the use of data mining techniques to find and obtain pertinent information (knowledge) from WWW documents and services. Web mining can be broken down into three categories: web usage mining, web structure mining, and web content mining [2, 3].

Mining the content of online pages is known as web content mining (WCM). It can be utilised on web sites directly or on search engine result pages. Database (DB) View and Information Retrieval (IR) View are two different viewpoints from which WCM can be distinguished.In the IR perspective, practically all studies employ a collection of words to represent unstructured text, while the HTML structure found inside the pages can be used to represent semi-structured data. Web mining can be done here using intelligent web agents. Web mining attempts to determine the structure of the website from a multi-level database that can be translated into the representation of a web site in DB perspective

Web Content Mining (WCM), which focuses on the structure of inner documents, Web Structure Mining (WSM) aims to identify the link structure of hyperlinks at the inter-document level. In a web graph, where web pages serve as nodes and hyperlinks as edges connecting related pages, it is used to construct structural summaries about the online pages.online Usage Mining (WUM) is a technique used to identify user travel patterns and meaningful information from online data kept in server logs as a result of user interaction when browsing the internet. Finding broad access patterns or patterns that fit the given parameters can be added to the categories.

Web Usage Mining (WUM) is a technique used to identify user travel patterns and meaningful information from online data kept in server logs as a result of user interaction when browsing the internet. Finding the general access patterns or patterns that match the given specifications can be further classified under this heading.

The three web mining categories mentioned above each have their own set of use cases, such as site optimisation, business intelligence, online personalisation, site modification, usage characterization and classification, page ranking, etc. Search engines typically employ page ranking to identify more essential pages Different page ranking algorithms have been reported in the available literature [4, 5, 6, 8, 9, 10]. In the next section, four important page ranking algorithms: PageRank, Weighted PageRank, HITS and Page Content Rank, have been discussed giving details of their working.

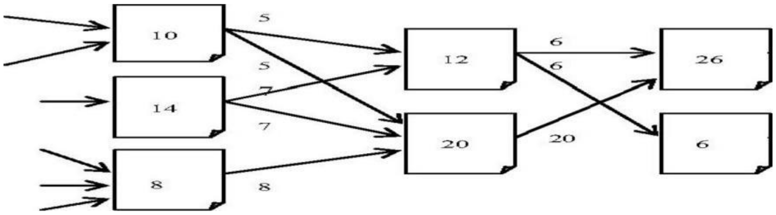
Page Rank Algorithm

The size of the WWW is expanding quickly, and at the same time, the volume of searches that search engines can process has increased dramatically. The quantity of search engine inquiries is growing dramatically along with the number of users on the internet. The search engine must therefore be able to process these inquiries quickly. In order to extract just pertinent documents from the database and give consumers the desired information, some sort of web mining approach must be used.

Page ranking algorithms, which can arrange the documents in order of their relevance, importance, and content score and use web mining techniques to sort them, are used to offer the documents in an ordered manner. Some algorithms use both links and the content of the document to assign a particular document, whereas others use only the link structure of the documents—their popularity scores—while still others look for the content in the documents (web content mining). The following list of common page ranking algorithms has been discussed.

PageRank algorithm, to start

PageRank (PR), named after Larry Page, a cofounder of the Google search engine, was created by Surgey Brin and Larry Page[5, 6]. It ranks websites based on their value using the web's link structure. Google[7] utilises PageRank to rank the search results in such a way that documents that appear to be more important get to the top of the list. According to this algorithm, if a page has some significant incoming links, then its outgoing links to other pages also become significant. Backlinks are considered as a result, and links are used to spread the ranking.As a result, a page receives a high rank if the total of its PageRank, Weighted PageRank, HITS, and Page Content is high.More than 25 billion web pages on the WWW are taken into account by the PageRank algorithm for determining a rank score [7]. To determine an overall ranking score for each returned web page in response to a query, Google blends precomputed PageRank scores with text matching scores [11]. Although a variety of factors are taken into account for calculating overall rank, Google's PageRank algorithm is at its core.



l) Example Illustrating Working ofPR

To explain the working of PageRank, let us take an example hyperlinked structure shown in Fig. 5, where A, B and C are three web pages.

The PageRanks for pages A, B and C can be calculated by

$$\begin{aligned}&PR(A)=(1-d)+d((PR(B)/2+PR(C)/1\ )\\&PR(B)=(1-d)+d(\ PR(A)/2+PR(C)/1\ )\\&PR(C)=(1-d)+d(\ PR(B)/2)\end{aligned} \qquad (2b)$$

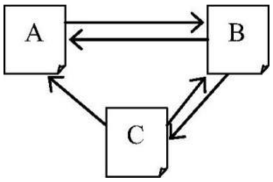


Figure 5. Example Hyperliked Structure

By calculating the above equations with (1—0.5 (say), the page ranks of pages A, B and C become:

$PR(A)=1.2, PR(B)=1.2, PR(C)=0.8$

) **Page Rank Iterative Method**

For a limited group of pages, it is simple to solve the equation system to obtain page rank values, but because the web has billions of documents, it is impossible to discover a solution using the inspection approach

TABLE I ITERATION METHOD OF PAGERANK

Iteration	PR(A)	PR(B)	PR(C)
	1	1	1
1	1	1.25	0.81
2	1.21	1.2	0.8
3	1.2	1.2	0.8
4	1.2	1.2	0.8

It may be noted that in this example, PR(A)=PR(B)>PR(C), Experiments have shown that rank value of a page converges to reasonable tolerance in roughly logarithmic (log n) [5, 6].

Weighted Page Rank Algorithm

Weighted PageRank (WPR) is an addition to conventional PageRank that was proposed by Wenpu Xing and Ali Ghorbani [10]. It is predicated on the idea that more popular websites will link to or be connected from more popular websites. This approach gives pages with higher importance higher rank values rather than dividing a page's rank value equally among its outbound connected pages. Each outbound link receives a value based on how well-known or significant it is. The quantity of inbound and outbound connections to a page serves as a gauge of its popularity. The incoming links are given a weight value that represents the popularity.The weight values used to give popularity to the inbound and outbound links are Wvt(v,u) and W9ut(v,u), respectively. The weight of the link (V,U) is W'1 (V,U), which is determined by the number of incoming links to page u and the number of incoming links to all reference (outgoing linked) sites on page v.

Weighted PageRank formula is given as:

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

Page Content Rank Algorithm

Page Content Rank (PCR) is a new ranking approach of page relevance ranking that Jaroslav Pokorny and Jozef Smizansky[4] presented. This approach incorporates several criteria that appear to be crucial for examining the content of web pages. Here, the importance of the terms on a page is used to establish the relevance of the page, and the importance of a term is described in relation to a specific query, q. A neural network serves as the PCR's internal categorization framework.

In PCR, assume that for a given query q and a typical search engine, a set Rq of ranked pages is the output, and these sites are then categorised according on their significance.Similar to the vector model [12], a page is represented here using frequencies of terms found on the page.

) Working ofPCR

PCR method can be described in the following four steps:

(i) Term extraction: An HTML parser extracts terms from each page in Rq. An inverted list [13] is built in this step which is used in step (iv).

(ii) Parameter Calculation: Statistical parameters such as a Term Frequency (TF) and occurrence positions; as well as linguistic parameters such as frequency of words in the natural language are calculated and synonym classes are identified.

(iii) Term classification: Based on parameter calculations in step (ii), the importance of each term is determined. A neural network is used as a classifier that is learnt on a training set of terms. Each parameter corresponds to excitation of one neuron in the input level and the importance of a term is given by excitation of the output neuron in the time of termination of propagation.

(iv) Relevance Calculation: Page relevance scores are determined on the basis of importance of terms in the page, which have been calculated in step (iii). The new score of a page P is equal to the average importance of terms in P.

PCR asserts that the importance of a page P is proportional to the importance of all terms in P. This algorithm uses the usual aggregation functions like Sum, Min, Max, Average, Count and also a function called Sec moment

PCR parameter estimation

A term's importance, denoted by importance(t), is calculated using 5+(2*NEIB) parameters, where NEIB stands for the number of neighbouring terms that were considered in the calculation. The computation depends on factors such database D, query q, and the total number of pages taken into account (n). Additionally, the classification function classify is utilised, yielding the importance of t for 5+(2*NEIB) parameters.

Page Classification and Significance Calculation

PCR uses neural network technique as a classification tool that is considered NET. NET sets weights from previous tests. Let's calculate a general neural network NET**,**

assuming 5 + (2 * NEIB) neurons in the input layer of the network and one neuron in the output layer, and the input vector v NET**,** denoted as NET(v) and NE II i]**.** output Firing

the ith neuron in the layer after the computation is complete. The classify() function can be

defined as:

Page importance is calculated as the importance of all objects in P, so

page importance (P) = seconds moment({importance(t): teP }

values per page ,

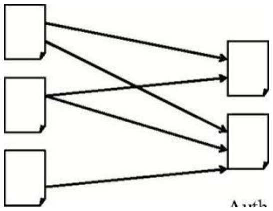
18) Rq n gives a new ranking to the first n pages in the search engine. This new

layout represents pages according to their content, not PR and WPR.

HITS Algorithm

For each query entered by the user, Kleinberg [8] created a WSM-based algorithm called Hyperlink-Induced Topic Search (HITS) [9] that makes the assumption that there are a number of authority pages that are pertinent and well-liked and focus on the query, as well as a number of hub pages that contain helpful links to relevant pages/sites, including links to many authorities.

HITS presupposes that if page p's author links to page q, page p must provide page q some sort of authority.The WWW is viewed by the HITS method as a directed graph G(V,E), where V is a collection of vertices that represent pages. A link from page p to page q is indicated by a directed edge (p, q).The first results that the search engine returns are an excellent place to start because it's possible that it didn't find all the pages that were relevant to the query. However, depending solely on the first few pages does not ensure that authority and hub pages are also efficiently retrieved. In order to solve this issue, HITS employs a suitable technique to locate the pertinent data pertaining to the user query.



Working of HITS

The first step in the HITS algorithm is to find the most relevant pages for the search query. This set is called the base set and can be obtained by fetching the first page returned from the text-based search algorithm. The base set is created by extending it with all linked web pages from the base set and some pages linked to it. The web pages in the base set and all hyperlinks on those pages create a focused subgraph. HITS count is only done on this map.The reason for creating a light base, according to Kleinberg, is to ensure that most (or most) of the strongest officers are included.   In mutual iteration, the rule and the base value are defined in relation to each other. The correct value is calculated as the sum of the zoom center values for that page. The average value is the number of right zoom values of the page it points to. Some implementations also take into account the effect of the contact page.The algorithm performs a series of iterations, each with two initial stages: Authority update: Update each node's authority score to be equal to the score of all points for itself. That is, nodes are given a high score by connecting through the pages of the data centers. Hub Update: Update each of the hub scores to be equal to the score of the rule at each point it points to. In other words, nodes get a high hub score by connecting to nodes that are considered authoritative on the subject. The hub score and policy score of nodes are calculated with the following algorithm: starts with one hub score and 1 policy score for each node.Run update rule Run update rule , normalize the value of the rule scores by dividing each center score by the square root of the sum of the squares of all scores in the center, and divide each score by the root of the sum of squares. total score center. Repeat from step 2 as needed.

The first step in the HITS algorithm is to find the most relevant pages for the search query. This set is called the base set and can be obtained by fetching the first page returned from the text-based search algorithm. The base set is created by extending it with all linked web pages from the base set and some pages linked to it. The web pages in the base set and all hyperlinks on those pages create a focused subgraph. HITS count is only done on this map.The reason for creating a light base, according to Kleinberg, is to ensure that most (or most) of the strongest officers are included. In mutual iteration, the rule and the base value are defined in relation to each other. The correct value is calculated as the sum of the zoom center values for that page. The average value is the number of right zoom values of the page it points to. Some implementations also take into account the effect of the contact page.The algorithm performs a series of iterations, each with two initial stages: Authority update: Update each node's authority score to be equal to the score of all points for itself. That is, nodes are given a high score by connecting through the pages of the data centers. Hub Update: Update each of the hub scores to be equal to the score of the rule at each point it points to. In other words, nodes get a high hub score by connecting to nodes that are considered authoritative on the subject. The hub score and policy score of nodes are calculated with the following algorithm: starts with one hub score and 1 policy score for each node.Run update rule Run update rule , normalize the value of the rule scores by dividing each center score by the square root of the sum of the squares of all scores in the center, and divide each score by the root of the sum of squares. total score center. Repeat from step 2 as needed.

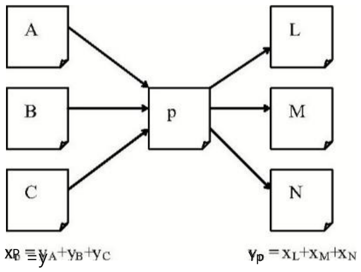


Figure. 9. An example of HITS operation

Comparative Research

A close examination of the existing literature reveals that various Hubs and authorities are given relative weights, with each algorithm using different fundamental ideas.Although HITS is an iterative method based on the link structure of online publications, like PageRank and WPR, it does have some significant peculiarities. It is processed on a calculated as follows: It is executed at query time, not at indexing time; it calculates two scores per document rather than one; Not all documents, but a tiny portion of those that are important.

Additionally, like HITS, PCR selects a selection of documents from the result list and uses a classification technique that is not included in HITS. With PCR and HITS, PR and WPR set themselves apart from the competition since they primarily concentrate on the hyperlink structure of the pages rather than their content.

TABLE II COWARISON OF PAGE RANKING ALGORITHMS

Algorithm	PageRank	Weighted PageRank	Page Content Rank	HITS
Main Technique Used	Web Structure Mining	Web Structure Mining	Web Content Mining	Web Structure Mining, Web content mining
Description	Computes scores at indexing time not at query time. Results are sorted according to importance of pages.	Computes scores at indexing time, unequal distribution of score, pages are sorted according to im ortance.	Computes new scores of the top n pages on the fly. Pages returned are related to the query i.e. relevant documents are returned.	Computes hub and authority scores of n highly relevant pages on the fly. Relevant as well as important pages are returned.
UP Parameters	B acklinks	Backlinks, forward links	Content	Backlinks, forward links, content
Working levels		1	1	
Complexity	O(log N)	< O(log N)		< O(log N) (higher than WPR)
Relevancy	Less	Less hi er than PR	More	More less than PCR
Importance	More	More	less	less
Quality of result	Medium	Hi her than PR	A rox e ual to WPR	Less than PR
Limitations	Computes scores at indexing time not on fly. Results are sorted according to importance of a es.	Relevancy is ignored. Method computes scores at a single level.	Importance of pages is totally ignored.	Topic drift and efficiency problems

CONCLUSION

Using web mining, it is possible to glean meaningful information from a vast amount of web data. The typical search engines typically return a lot of pages in answer to users' queries, but users always want the best results quickly, so they don't bother to browse through all the pages to find the ones they need. The page ranking algorithms, a web mining programme, are crucial in facilitating user search navigation inside search engine results.The PageRank and Weighted Page Rank algorithms prioritise links over page content, the HITS algorithm emphasises both page content and links, and the Page Content Rank algorithm simply takes into account page content.

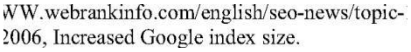
REFRENCES

Companion slides Ibr the text by Dr. M. H.Dunham,

Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An Approach to the Web Content Mining".

L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pageranl< Citation Ranking: Bringing order to the Web". Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

C. Ridings and M. Shishigin, "Pagerank Uncovered". Technical report,
2002.


http : WW.webrankinfo.com/english/seo-news/topic-
2006, Increased Google index size. //WWW. 1 63 88. htm.January

Kleinberg J. , "Authorative Sources in a llyperlinked Environment".
Proceedings of the 23rd annual International ACM SIGIR Confereon Research and Development in InFörmation Retrieval, 1998.

C. I)ing, X. I le, P. I lusbands, I I. Zha, and I I. Simon, "Link Analysis:
I lubs and Authorities on the World". Technical report:47847,2001.