Comparative Study of Different Page Rank Algorithms

Chintan Agravat
LD college of engineering, LDCE, Ahmedabad, India

Abstract: **web pages are increasing day by day. Due to that problem arises how to give better search result for a given search engine query. Here I have shown a comparison between various page rank algorithms used for displaying Quality results for various search engine queries.**

Index Terms— **Ranker, WUM ,WSM ,WCM,QI.**

I. INTRODUCTION

Here the basic search engine architecture is shown from the given architecture. We clearly see that how the Search engine uses different steps to execute query. Here as shown in figure ranker is there. It uses different ranking algorithm For Quality Results. An efficient ranking of query words has a major role in efficient searching for query words[2].

Ranker to calculate scores for the matched documents according to their relevancies to the input query (L1 ranking), and select the top *k* Web pages to recalculate the scores with a more complex function (L2 ranking). The ranking functions are usually the combination of many Information Retrieval (IR) document features, which includes some query dependent dynamic and some query independent static features[1]. page rank is static query independent.

In Section II -related work, III –comparison of ranking algorithms,IV-conclusion.

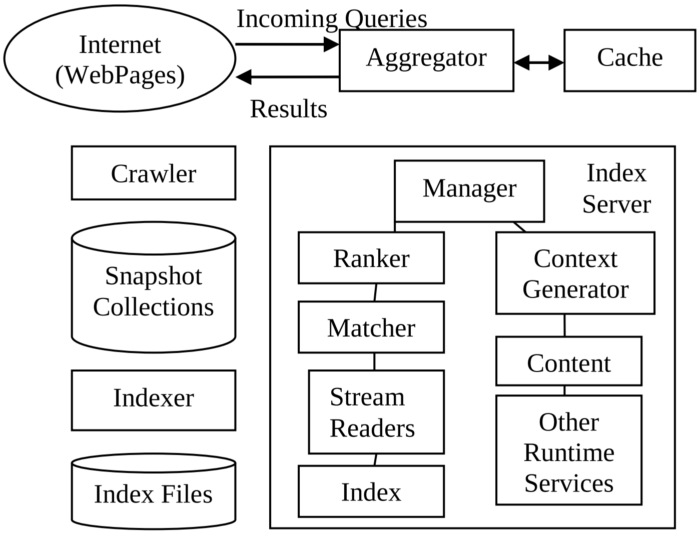

Fig1:Basic Search Engine Architecture[1].

II. RELATED WORK

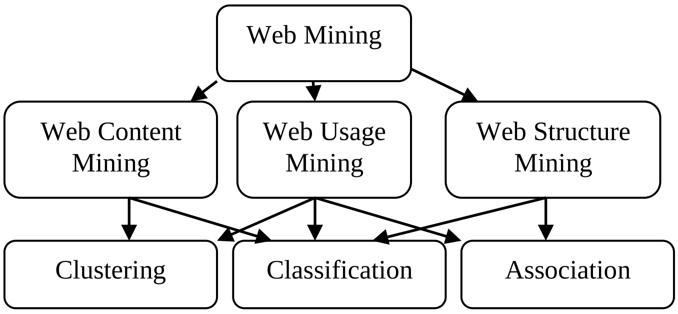

Fig2: Web Mining Categories.

Web Content Mining: Web content mining deals with extracting information from the content data that is present in the web page to be conveyed to the user. The content of web pages can be text, image, audio, video, etc[3].

Web Structure Mining: Web Structure Mining is the process of discovering Structure information in the web. This is used to analyse the link Structure of the web it uses two structures *hyperlink* and *Document*[3].

Web Usage Mining: Web Usage Mining is the process of discovering information from the usage data of the web pages.

Different variations of page rank algorithm falls under different categories of web mining. According to methods used to construct algorithm. It may use combinations of web mining methods. Generally these algorithm falls under web structure mining. As link is main concern in these algorithms.

A. *Page Rank Algorithm*

Page and Brin proposed a formula to calculate the PageRank of a page A as stated below-

$$PR(A)=(1d)+d(PR(T1)/C(T1)+\ldots..+PR(Tn/C(Tn))$$

here PR(Ti) is the PageRank of the Pages Ti which links to page A, C(Ti) is number of outlinks on page Ti and d is damping factor. It is used to stop other pages having too much influence. The total vote is “damped down” by multiplying it to 0.85[5].

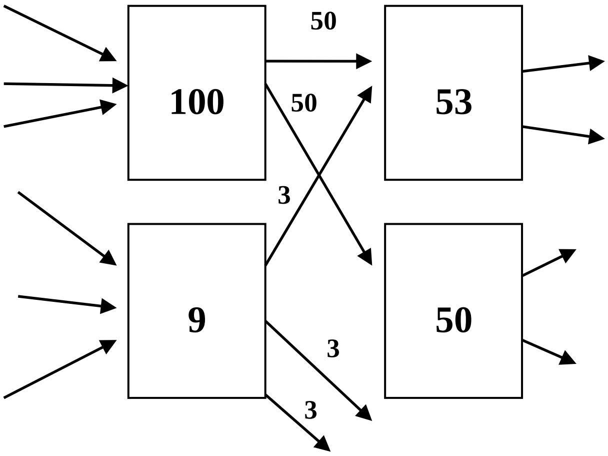

Fig 3: Simplified Calculation of Page Rank.

There is a little problem with so called rank sinks. Those are *closed loops* of pages that accumulate rank but never distribute it further [5]. Page rank algorithm has complexity of Log n [5].

Limitations

- PageRank is equally distributed to outgoing links[7].
- It is purely based on the number of in-links and out-links[7].

B. Weighted Page Rank Algorithm

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm. This Algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links. This is denoted as **Win (m,n)** and **Wout (m,n)** respectively.

Win (m,n) is the weight of link(m,n) . It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$W_{(m,n)}^{in} = \frac{I_n}{\sum_{p\in R(m)} I_p}$$

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m. **Wout (m,n)** is the weight of link(m,n). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W_{(m,n)}^{out} = \frac{O_n}{\sum_{p\in R(m)} O_p}$$

On is number of outgoing links of page n, Op is number of outgoing links of page p, Then the weighted PageRank is given by formula:

$$WPR(n) = (1-d) + d \sum_{m\in B(n)} WPR(m)\, w_{(m,n)}^{out} w_{(m,n)}^{in}$$

The complexity of Page Rank Algorithm is < log n[5].

Limitations

- Certain pages which are not relevant to the query are also included in the result set because of their popularity. For example home pages are also ranked higher because of their greater number of inlinks and outlinks
- Topic Drift can occur.
- Weightage distributed proportionately to pages based on popularity.
- Query Independent[3].

C. HITS (Hyper-link Induced Topic Search)

Klienberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time. The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 3 shows the hubs and authorities in web.

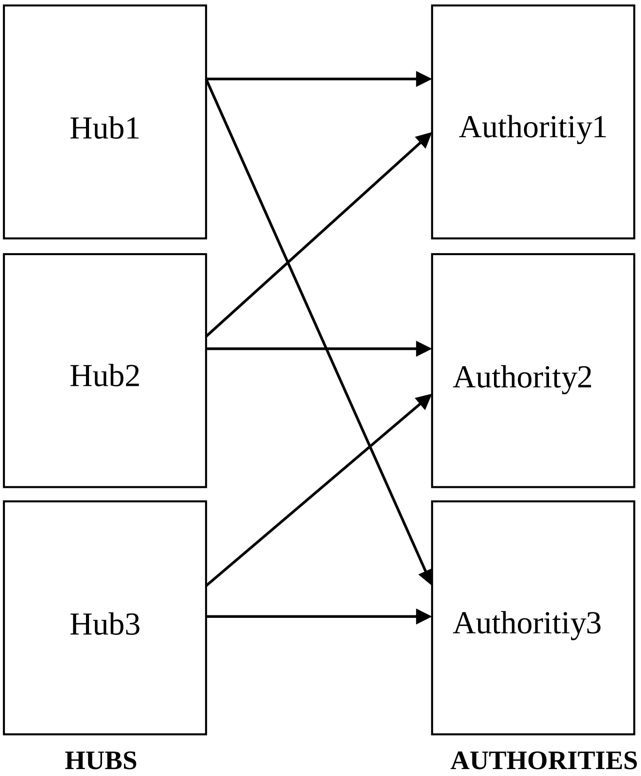

Fig 4:Hubs and authority in HITS.

In It has two steps:

1. Sampling Step: In this step a set of relevant pages for the given query are collected.

2. Iterative Step: In this step Hubs and Authorities are found using the output of sampling step.

Following expressions are used to calculate the weight of

Hub (Hp) and the weight of Authority (Ap).

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

here Hq is Hub Score of a page, Aq is authority score of a page, I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

The complexity of HITS algorithm is < log n [5].

Limitations

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query.
- Efficiency: HITS algorithm is not efficient in real time. HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints HITS could not be implemented in a real time search engine.

D. Query Dependent PageRank

The Query Dependent page model uses the intelligent surfer model. This query dependent page rank attempts to rank pages based upon the query. Initially the relevance of each page to the query term is found out. Then the PageRank is calculated based upon the Relevance Measure.

The intelligent surfer Model: The intelligent surfer algorithm attempts to improve upon the standard PageRank algorithm by introducing a more intelligent random surfer . This surfer is guided by a probabilistic model of relevance to the query, with the probability distribution given by:

$$P_q(j) = (1-d)P'_q(j) + d \sum_{t \in B(j)} P_q(t)\, P_q(t \to j)$$

Where *Pq(j)* is the PageRank of page ‘*j*’ for query *q*

d is the damping factor (0-1)

Pq(i-> j) is the probability that the surfer goes to page ‘*j*’ from page ‘*i*

The probability that the surfer jumps when not following links is specified by *(1-d)Pq’(j)*. The resulting probability distribution over pages is given by *Pq(j)* and both *Pq’(j)* and *Pq(i→ j)* are derived from a measure of relevance of page *j* to query *q*, and are given by:

$$P'_q(j) = \frac{R_q(j)}{\sum_{k \in w} R_q(k)}$$

$$P_q(t \to j) = \frac{R_q(j)}{\sum_{k \epsilon F_t} R_q(k)}$$

Where *Rq(j)* is the relevance of the page *j* to the query q.

The web surfer probabilistically hops from page to page, choosing the pages that seem relevant to the query. This intelligent web surfer model proposes a query-dependant PageRank in which the score depends not only on the number of backlinks, but also on the relevance of the query to that page. This model will provide a higher quality of PageRank compared to the random surfer model. When given a query of multiple terms one term is selected based upon some probability distribution and it is used to guide the behaviour of other terms in the query.

E. SQD PageRank

SQD PageRank is another improvement of Query Dependent PageRank algorithm [16]. The SQD PageRank is used for multi-term queries. This is simultaneous multiple term Query Dependent PageRank. The SQD –PageRank is most suitable for queries with many terms in it. This algorithm measures the relevance of each term in the query to the entire set of web pages. Let us consider a query of multiple terms given by Q={q1,q2,q3,.....qn}. The weight of each term in the query is given by {k1,k2,k3......kn}. *Rqi (j)* denotes the relevance of page *j* to the query term *qi*. The Relevance of the page *j* to the query *q* is given by

$$R_{\otimes_{i=1}^{n} q_i}(j) = \frac{\sum_{i=1}^{n} k_i * R_{q_i}(j)}{\sum_{i=1}^{n} k_i}$$

$$P_{\otimes_{i=1}^{n} q_i}(j) = (1-d)P^t_{\otimes_{i=1}^{n} q_i}(j) + d \sum_{t \in B_j} P_{\otimes_{i=1}^{n} q_i}(t) P_{\otimes_{i=1}^{n}}$$

$$P^t_{\otimes_{i=1}^{n} q_i}(j) = \frac{R_{\otimes_{i=1}^{n} q_i}(j)}{\sum_{k \in w} R_{\otimes_{i=1}^{n} q_i}(k)}$$

$$P_q(t \to j) = \frac{R_{\otimes_{i=1}^{n} q_i}(j)}{\sum_{k \in F_t} R_{\otimes_{i=1}^{n} q_i}(k)}$$

This formula finds the relevance of all the query terms to all the web pages in the web graph, the SQD page rank algorithm has also been improved based upon the statistical and semantic relevance between pages. It includes

Ontology into SQD pageRank which is termed as ONTO-SQD pageRank.

F. DistanceRank

Distance rank is an intelligent ranking algorithm that is based upon reinforcement learning .In this ranking method, the distance between pages is considered. The distance is given by the number of links between the pages. That is the distance between page *i* and page *j* is dependent on how many links that the surfer has to traverse in order to reach page *j* from page *i*. The distance between pages is taken as a penalty and the pages longer links have lesser ranks. This distance rank also follows the properties of PageRank algorithms in which a page gets a higher rank if it has many links, And if it is linked by a page with a higher rank. Similarly in distance rank algorithm a page gets a higher rank if it has many links thereby having a shortest link, And if a page with small distance point to this page.

$$D_b = \frac{\sum_{t=1}^{v} D_{ab}}{V}$$

Where V is the number of pages present in the web[3].
db is the average distance of the page b in the web[3].

G. Weighted Links Rank Algorithm

A modification of the standard page rank algorithm is given by Ricardo Baeza-Yates and Emilio Davis named as weighted links rank (WLRank). This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are improved using weighted links. The length of anchor text seems to be the best attributes in this algorithm. Relative position, which reveal that physical position does not always in synchronism with logical position is not so result oriented. Future work in this algorithm includes, tuning of the weight factor of every term for further evolution[2].

H. EigenRumor Algorithm

As the number of blogging sites is increasing day by day, there is a challenge for service provider to provide good blogs to the users. Page rank and HITS are very promising in providing the rank value to the blogs but some limitations arise, if these two algorithms are applied directly to the blogs The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, a EigenRumor algorithm is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of eigen vector[3].

I. Time Rank Algorithm

An algorithm named as TimeRank, for improving the rank score by using the visit time of the web page is proposed by H Jiang et al. Authors have measured the visit time of the page after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm[2].

J. TagRank Algorithm

A novel algorithm named as TagRank for ranking the web page based on social annotations is proposed by Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun. This algorithm calculates the heat of the tags by using time factor of the new data source tag and the annotations behaviour of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very accurate and this algorithm index new information resources in a better way. Future work in this direction can be to utilize co-occurrence factor of the tag to determine weight of the tag and this algorithm can also be improved by using semantic relationship among the co-occurrence tags[2].

K. Relation Based Algorithm

Fabrizio Lamberti, Andrea Sanna and Claudio Demartini proposed a relation based algorithm for the ranking the web page for semantic web search engine. Various search engines are presented for better information extraction by using relations of the semantic web. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Results are very encouraging on the parameter of time complexity and accuracy. Further improvement in this algorithm can be the increased use of scalability into future semantic web repositories[2].

III. COMPARISON OF VARIOUS WEB PAGE RANKING ALGORITHMS

*H*ere I have ceated a table showing comparision of various PageRank Algorithms Explained above and also I have shown comparision of different parameters that how one algorithm outperform others and limitations of algorithms as compared to other algorithms. This table can give us outline that which algorithm is better for the particular type of surfer.

Algorithms	PageRank	HITS	Weighted PageRank	EigenRumor
Main Technique	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Content Mining	Web Structure Mining, Web Content Mining
Methodology	This algorithm computes the score for pages at the time of indexing of the pages.	It computes the hubs and authority of the Relevant pages. It relevant as well as important page as the Result.	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of Page is decided.	Eigenrumor use the adjacency matrix, which is constructed from agent to object link not page to page link
Input Parameter	Back links	Content, Back and Forward links	Back links	Agent/Object
Relevancy	Less (this algo. rank the pages on the indexing time)	More (this algo. Uses the hyperlinks so according to Henzinger, 2001 it will give good results and also consider the content of the page)	Less as ranking is based on the calculation of weight of the web page at the time of indexing.	High for Blog so it is mainly used for blog ranking.
Quality of results	Medium	Less than PR	Higher than PR	Higher than PR and HITS
Importance	High. Back links are considered.	Moderate. Hub & authorities scores are utilized.	High. The pages are sorted according to the importance.	High for blog ranking.
Limitation	Results come at the time of indexing and not at the query time	Topic drift and efficiency problem	Relevancy is ignored.	It is most specifically used for blog ranking not for web page ranking as other ranking like page rank, HITS.

Algorithms	Weighted Link Rank	Distance Rank	Time Rank	Tag Rank
Main Technique	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Usages Mining	Web Content Mining
Methodology	it gives different weight to web links based on 3 attributes: Relative position in page, tag where link is contained, length of anchor text.	Based on reinforcement learning which consider the logarithmic distance between the pages.	In this algorithm the visiting time is added to the computational score of the original page rank of that page	Visitor time is used for ranking. Use of sequential clicking for sequence vector calculation with the uses of random surfing model.

Input Parameter	Content, Back and Forward links	Forward links	Original Page Rank and Sever Log	Popular tags and related bookmarks
Relevancy	more (it consider the relative position of the pages)	Moderate due to the use of the hyperlinks.	High due to the updation of the original rank according to the visitor time.	Less as it uses the keyword entered by the user and match with the page title.
Quality of results	Medium	High	Moderate	Less
Importance	Not specifically quoted.	High. It is based on distance between the pages.	High, Consideration of the most recently visited pages .	High for social site
Limitation	Relative position was not so effective, indicating that the logical position not always matches the physical position.	If new page inserted between two pages then the crawler should perform a large calculation to calculate the distance vector.	Important pages are ignored because it increases the rank of those web pages which are opened for long time.	It is comparison based approach so it requires more site as input.

Algorithms	**SQD Pagerank**	**Relational Based Page Rank**	**Query Dependent Ranking**
Main Technique	Web Content Mining	Web Structure Mining	Web Content Mining
Methodology	Measures relevance for all term in query & computes PageRank	A semantic search engine would take into account keywords and would return page only if both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.	This paper proposed the construction of the rank model by combining the results of similar type queries
Input Parameter	Query & Inlinks	Keywords	Training query
Relevancy	High(because it uses intelligent surfer model)	High as it is keyword based algorithm so it only returns the result if the keyword entered by the user match with the page.	High (because the model is constructed from the training quires).

Quality of results	High	High	High
Importance	Moderate(Query dependent)	High. Keyword based searching.	High because it gives the results for user's query as well as results for similar type of query.
Limitation	Identifying important terms in the query	In this ranking algorithm every page is to be annotated with respect to some ontology, which is the very tough task.	Limited number of characteristics are used to calculate the similarity.

IV.CONCLUSION

With the survey of these different Search engine Algorithms we can conclude that some algorithms give more relevant results but they are time consuming and some algorithms have moderate relevancy in the result but they are time efficient.

REFERENCES

[1]. Jing YAN1, Zhan-Xiang ZHAO1,2, Ning-Yi XU1, Xi JIN2, Lin-Tao ZHANG1, Feng-Hsiung HSU1 ,1Microsoft Research Asia, Beijing, China,

2University of Science and Technology of China, Hefei, China,

[2]. Dilip Kumar Sharma ,GLA University, Mathura, UP, India,todilipsharma@rediffmail.com

A. K. Sharma,

YMCA University of Science and Technology,

Faridabad, Haryana, India.

[3]. Ms.M.Sangeetha,

Department of Computer Science,

School of Engineering and Technology, Pondicherry University,

Puducherry, India

sangeepower@gmail.com

Dr.K.Suresh Joseph,

Department of Computer Science,

School of Engineering and Technology, Pondicherry University,

Puducherry, India,

ksjoseph.csc@gmail.com

Page-

[4]. Lissa Rodrigues

Computer Engineering

St. Francis Institute ofTechnology

Mumbai, Maharashtra

lis 10rods@gmail.com.

Shree Jaswal

Information Technology

st. Francis Institute of Technology

Mumbai, Maharashtra

shreejani@gmail.com.

Page-

[5]. Ashish Jain, Rajeev Sharma, Gireesh Dixit

Madhav Proudyigiki Mahavidyalaya, Bhopal

ajpost@rediffmail.com, sharmaraj2007@gmail.com,

gireeshdixit15@rediffmail.com

Varsha Tomar

Jiwaji University, Gwalior

tomar.varsha86@gmail.com

[6].Nagappan, V.K 1,Dr. P. Elango2

1Research Scholar, Research and Development Center

Bharathiar University, Coimbatore

2Assistant Professor, Dept. of Informationl Technology,