Survey on Web Page Ranking Algorithms

Mercy Paul Selvan
M.E, Department of Computer Scienc Sathyabama University

A .Chandra Sekar
M.E Ph.D,Department Of Computer Science
St.Joseph's College of Engineering

A.Priya Dharshin
Department of Electronic Science
Sathyabama University

ABSTRACT:

Day by day the growth of the World Wide Web is increasing very rapidly. One of the research quotes "There are more than 11.3 billion web pages in the World Wide Web". Search engines help the user to surf the web. Due to the vast number of web pages it is highly impossible for the user to retrieve the apt web page he needs. Thus, Web search ranking algorithms play an important role in ranking web pages so that the user could retrieve the page which is most relevant to the user's query. This paper presents a study of some useful web Page ranking algorithms and comparison of these algorithms.

Keywords

Web, Search engine, Page ranking.

1. INTRODUCTION:

WWW is a vast resource of hyperlinked and heterogeneous information including text, audio, video and metadata. It is estimated that WWW is expanded by about 2000% since its evolution and is doubling in size every six to ten months [1]. Due to the rapid growth of information resources in WWW it is difficult to manage the information on the web. Therefore it has increasingly necessary for the users to use efficient information retrieval techniques to find and order the desired information. Search engines play an important role in searching web pages. The search engines gather, analyze, organize and handle the data from the internet and offers the users an interface to retrieval the network resources [2]. But the search engines returns thousands of results which includes a mixture of relevant and irrelevant information [3]. It is true that nearly 65% - 70% users will choose the first page of the return results and about 20% - 25% may choose the second page and very few of 3% - 4% users only check the remaining results [4]. This means that search engines must return good results which can satisfy the user's interest. Fig.1 shows the concept of Search engines. Search engines are used to find information from the WWW. They download, index and store hundreds of millions of web pages. They answer millions of queries every day. They act like content aggregators as they keep record of every information available on the WWW [5].

Fig 1: Concept of Search Engine

Web search ranking algorithms play an important role in ranking web pages so that the user could get the good result which is more relevant to the user's query.

In this paper, the survey of different web page ranking algorithms and comparisons of these algorithms are carried out. The structure of this paper is as follows: section 2 provides an overview of some important web page ranking algorithms, section 3 discusses the comparisons of these algorithms and finally in section 4 the paper is concluded.

2. WEB PAGE RANKING ALGORITHMS

The size of the www is growing rapidly and at the same time the number of queries can handle has also grown incredibly. With increasing number of users on the web, the number of queries submitted to the search engines are also growing exponentially. Therefore the search engines must be able to process these queries efficiently. Thus the web mining techniques are applied in order to extract only relevant documents from the database and provide intended information to the users.

To present the documents in an ordered manner, web page ranking methods are applied which can arrange the documents in order of their relevance, importance and content score and use web mining techniques to order them [5].

The ranking algorithms can be proposed on different categories and some of them are

- Link analysis algorithm
- Personalized web search ranking algorithms
- Page Segmentation algorithms

2.1 Link Analysis Algorithm:

The link analysis algorithm is based on the links structure of the documents. The quality of results from search engines is generally lower than what the user expects and this quality can be improved greatly if pages are ranked according to some criteria based on links between the pages., i.e. a page which has many references must have something to say.

2.1.1 Page rank Algorithm:

Surgey Brin and L.Page [6] proposed an algorithm called page rank algorithm. This algorithm is used by Google to rank the web pages. The PageRank results from a mathematical algorithm based on the web graph, (Fig. 2 shows the principle of PageRank algorithm) i.e. the web pages as nodes and links as edges. Rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page depends on the number of links it has. A page that is linked to by many pages with high PageRank

receives a high rank itself. If there are no links to a web page then there is no support for that page.

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. A probability is expressed as a numeric value between 0 and 1. The PageRank value for any page **u** can be expressed as:

$$PR(u) = \sum_{v \epsilon B_u} \frac{PR(v)}{L(v)}$$

Fig 2: Principles of PageRank Algorithm

2.1.2HITS Algorithm:

Jon Kleinberg introduced Hyperlink-Induced Topic Search (HITS) [7] (also known as hubs and authorities) is a link analyses algorithm that rates Web pages. The hubs are served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs [16]. Fig.3 shows the hubs and authorities.

This scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

Authority Update Rule:

∀p, we update auth (p) to be:

$$\sum_{i=1}^{n} hub(i)$$

Where n is the total number of pages connected to p and i is a page connected to p. That is, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

Hub Update Rule:

∀p, we update hub (p) to be:

$$\sum_{i=1}^{n} auth(i)$$

Where n is the total number of pages p connects to and i is a page which p connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking pages.

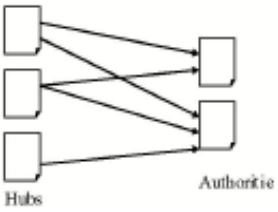

Fig 3: Hubs and Authorities

2.1.3Focused Rank:

Focused rank [8] is a link-based ranking algorithm based on the model of focused web surfers. Focused search points the interest of the particular topic. Focused search addresses the multi-dimensionality of web content and the relationship between pages of similar content.

If there is a hyperlink between page u and page v and they belong to at least one common topic, then link between these pages exist. . For a group of documents *S and* a list of topics τ we obtain a set of probabilities $P(d, ti)$, where $d \in S$ and $ti \in \tau$,of the document *d* belonging to a particular topic *ti*. The topical overlap between two documents is less when they have few topics in common. And also when two pages comes under same topic, the probability of their inclusion in those topics $P(d, ti)$ is also low, and therefore the topical overlap remains low.

The topical overlap score for a link exists from page *u* to page *v*, is computed for as

$$T(u, v) = \sum_{j \epsilon \tau} C_u(j)\, C_v(j)$$

The probability of the surfer navigating from page u to *v* is a function of the two pages topic overlap. Computing *T* for each pair of hyperlinked documents forms a matrix *M* where

$$Muv = T(u, v)$$

The probability of a focused surfer following a link from u to v is a function of the topical overlap scores between u and each of the pages to which it links

$$P_T(u \rightarrow v) = \frac{T(U,V)}{\sum_{d \epsilon D} T(u,d)}$$

where $T(u, d)$ is 0 if *u* does not link to *d*. $P_T(u \rightarrow v)$ is the portion of page *u*'s rank conferred to page *v*. Since *T* is a function of the topics shared between two pages, P_T is a model of focused surfing behavior as a function of the set of topics shared, the probability of inclusion in those topics, and the topology of the link graph.

The rank at iteration *i* of the link based ranking algorithm is

$$Rank_i = \frac{\alpha}{N} + (1- \alpha) \sum_{d \epsilon D} Rank_{i-1}(d)\; P_T(d \rightarrow u)$$

Where α is a scaling factor, N is the number of documents in D, and PT ($d \rightarrow u$) is equal to 0 if d does not link to u.

2.2 Personalized Web Search Ranking Algorithm:

The huge amount of information available on the internet is widely shared due to the ability of web search engines to find useful information for users. But the search engines return results which are definitely useless for the user. This is mainly due to the fact that they return results based on simple keyword matches without any concern for the information needs of the user at a particular instance of time.

Personalization is the process of gathering the experience of individual user. The main objective of the personalization is to deliver the user the more satisfied results by giving most relevant information [9]. During search, personalization involves the below steps

i) According to the user's interest, collect and represent the information about the user.

ii) Use the collected and represented information to re-rank the results returned from the initial retrieval process or directly includes the information into the search process itself to select personalized results.

2.2.1 An Integrated Page Ranking Algorithm:

Search engines use the history of user's interest and return results. But sometimes this is not sufficient. So in this paper, search results are ranked based on user preferences in content and link. The preference of the content and the link is integrated in order to rank the results.

[10]Page links and contents are used individually for the ranking process. Semantic relationships and hyperlink relations are used to improve the ranking mechanism. The semantic relationships indicate the content relevancy. The semantic relations are used to analyze the contents of the web pages. The ontology is used to analyze the content relationship. The hyperlink network is used to represent content referenced by the other pages. The hyperlink also shows the related sources.

The link based ranking scheme does not considers the content relationship. The content based ranking scheme does not consider the page content values.

The search query values are prepared using the semantic information. Domain selection is used to support the search query optimization. The result page analysis operations are performed under the client environment. The cleaning process is used to remove noisy data under the web pages. The results are ranked and irrelevant pages are removed from the result.

Link relevant is prioritized based on term weight. Term weight is assigned based on their search history weight of the user. Based on the user's preferences the relevant links will be prioritized. Rocchio is a relevance feedback method.

$$M(I, J) = \text{Max}\left(\sum_{K=1}^{m} DT(k, j) * DC(k, i)\right)$$

Where M is the matrix representing the user profile, I is the number of documents that are related to the i th category, m is the number of documents in DT, DT (k, j) is the weight of the j th term in the k th document, DC (k, j) is a binary value denoting whether the k th document is related to the i th category. Clearly, M (I, j) is the max weight of the j th term in all documents that are related to the I th category and documents that are not related to the category are not contributing to M (i, j).This method is called the Mrocchio method. Based on the category term weight, category link will be prioritized. To calculate category term weight, personalization for every user is included.

$$MU(i, j) = \sum_{u=1}^{n} \sum_{K=1}^{m} DT(k, j) * DC(k, i))$$

Interested term links will be mapped with the user and will be displayed.

It is shown in Fig.4 that precision is improved after integrated approach.

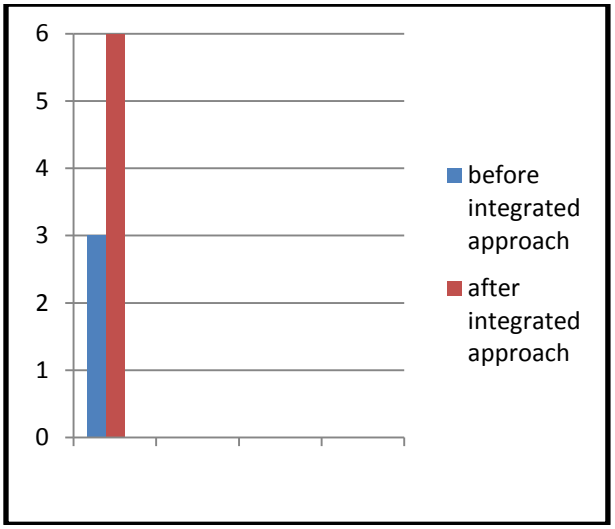

Fig 4: Comparison between Precision before And after Integrated Approach

Thus the ranking scheme is improved with the content and hyperlink in the web pages.

2.3 Page Segmentation Algorithms:

Page segmentation algorithm is to partition web pages into blocks. Because of the special characteristics of web pages like multiple topics and varying length of the web page, different page segmentation methods have impact on the web search performance. There are four main types of methods of page segmentation algorithms. They are

- Fixed-length Page segmentation
- DOM-based page segmentation
- Vision-based page segmentation
- Combined approach segmentation

2.3.1Fixed –Length Page Segmentation (FixedPS):

A fixed length passage contains fixed number of words. For web documents, fixed-length page segmentation is identical to traditional window approach except that all the HTML tags and attributes are removed. The only parameter is the length

of window and from the previous experience, it can be of 200 or 250. [11]

2.3.2DOM-Based Page Segmentation (DOMPS):

DOM provides each web page with a fine-grained structure, which explains not only the content but also the presentation of the page. In general, similar to discourse passages, the blocks produced by DOM-based methods tend to partition pages based on their pre-defined syntactic structure, i.e., the HTML tags.

2.3.3 Vision-Based Page Segmentation (VIPS):

People view a web page through a web browser and get a 2-D presentation which provides many visual cues to help distinguish different parts of the page, such as lines, images, etc [12].For the sake of easy browsing; a block within the web page is much likely about a single semantic. VIPS [13] is proposed to achieve a more accurate content structure on the semantic level.

2.3.4Combined Approach (CombPS):

It takes the advantage of both the visual layout and length normalization. In this, a web page will be first passed to VIPS for segmentation, and then to a normalization procedure.

Passage Retrieval:

Passage retrieval helps to apply retrieval algorithms to portions of a document, especially when documents have varying length. In passage retrieval, passages can be of three classes.

Discourse passages – rely on the logical structure of the documents marked by punctuation.[11]

Semantic passages - Obtained by partitioning a document into topics or sub topics according to its semantic structure.

Window passages – It contains fixed number of words. [11]

A natural correspondence between the page segmentation methods and traditional passage retrieval methods is shown in Table 1.

Table 1.Correspondence between Page Segmentation Methods and Traditional Passage Retrieval Methods

Webpage segmentation	FixedPS	DomPS	VIPS	CombPS
Passage Retrieval	Window	Discourse	Semantic	Semantic window

2.3.5 X-Y Cut Segmentation Algorithm:

The x-y cut segmentation algorithm [14], also referred to as recursive x-y cuts (RXYC) algorithm, and is a tree-based top-down algorithm. The root of the tree represents the entire document page. All the leaf nodes together represent the final segmentation. The RXYC algorithm recursively splits the document into two or more smaller rectangular zones which represent the nodes of the tree. At each step of the recursion, the horizontal and vertical projection profiles of each node are computed.

2.3.6 Voronoi – Diagram Based Algorithm:

The Voronoi-diagram based segmentation algorithm by Kiseet al. [15] is also a bottom-up algorithm. In the first step, it extracts sample points from the boundaries of the connected components using sampling rate rs. Then, noise removal is done using a maximum noise zone size threshold tn, in addition to width, height, and aspect ratio thresholds. After that a Voronoi diagram is generated using sample points obtained from the borders of the connected components. The Voronoi edges that pass through a connected component are deleted to obtain an area Voronoi diagram. Finally, superfluous Voronoi edges are deleted to obtain boundaries of document components.

3. COMPARISONS OF DIFFERENT ALGORITHMS

Each and every algorithm has its own advantages and disadvantages. The algorithm which comes under the different categories is compared under some parameters like basic criteria, merits, demerits, performance and relevancy. The table 2 and table 3 show the comparison of different algorithms based on link analysis and personalization web search, and page segmentation algorithms respectively.

Table 2. Comparisons of Algorithms Based on Link Analysis and Personalization Web Search

ALGORITHM	BASIC CRITERIA	MAIN TECHNIQUE	MERITS	DEMERITS	PERFORMANCE	RELEVANCY
PageRank	Link analysis algorithm based on random surfer model.	Web structure mining	Used in journal citations and in academic departments. Used to perform word sense disambiguation. The PageRank may also be used as a methodology to measure the apparent impact of a community like the blogosphere on the overall Web itself.	The main disadvantage is that it favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.	Medium	Less
HITS	Link analysis algorithm	Web structure and web content mining	Hub and Authority values are calculated so that the relevant and important pages are obtained.	Topic drift and efficiency problems occur. Non-relevant documents can be retrieved.	Less than page rank	More
Focused Rank	Link analysis based approach for focused surfer.	Web structure and web content mining	Good accuracy and relevancy is achieved without the online processing overhead.	It limits the scope of the documents indexed to those whose content falls under a particular topic	More	More
Integrated Page Ranking Algorithm	Page links and contents	Web structure, web content and web usage mining.	Irrelevant pages are discarded by post processing. Vocabulary problem is solved. Result refinement becomes easy. Noise data can be removed.	The page which is not linked by the user is considered as irrelevant page.	Medium	More

Table 3.Comparison of Page Segmentation Algorithms

ALGORITHM	BASIC CRITERIA	MAIN TECHNIQUE	MERITS	DEMERITS	RETRIEVAL PERFORMANCE
Fixed-Length Page Segmentation Algorithm	Fixed number of words or fixed length passages	Web Content Mining	Simplicity, Very robust, Effective For improving performance	No semantic information is taken into account in the segmentation process.	Medium
DOM-Based Page Segmentation Algorithm	Tags or tag types and also Content and link.	Web Content Mining	Provides a hierarchical structure of every web page.	Difficult to evaluate and compare.	Less
Vision Based Page Segmentation Algorithm	Visual cues	Web Content Mining	Achieve more accurate content structure on the semantic level. Greatly improve the performance of pseudo relevance feedback	Suffer from lack of normalization. It remains unclear.	Medium
Combined Approach Segmentation Algorithm	Visual layout and fixed length.	Web Content Mining	Advantage of both Visual layout and length normalization.	Little time consuming.	More
X-Y Cut	Top-down approach	Web Content Mining	It is Fast and easy to implement	Presence of noise Under segmentation errors	Medium
Voronoi Diagram Based Algorithm	Bottom up approach	Web Content Mining	Voronoi diagram based algorithm is good in Layouts having different variations	Spacing variations can occur. Over segmentation errors is introduced.	Medium

4. CONCLUSION

The algorithms that are described above are effective in retrieving the web pages from the search engines. The link analysis algorithms are based on link structure of the documents. The page which has many links has many references can improves retrieval efficiency. In the integrated ranking approach comes under personalized web search. In integrated approach both the content and the link are ntegrated to improve the retrieval efficiency. Page Segmentation algorithms are used to segment the page as blocks and by separating as blocks the retrieval performance in the web context could be improved. Each and every algorithm has got its own merits and demerits. As per the requirements of a search engine we can utilize the above said algorithms. It helps to enhance the current page rank algorithm used by the Google and these web page ranking algorithms could be used by several other search engines to improve the retrieval efficiency of the web pages as per the user's query.

REFERENCES

[1] Naresh barsagade, "Web Usage mining and Pattern Discovery: A Survey paper", CSE 8331, Dec.8, 2003.

[2] Justin Zobel, Alistair Moffat, Ron Sacks-Davis. "An Efficient indexing technique for full-text database systems".[c]//Proc of 18th Int Conf on VLDB, August 23 27, 1992, Vancouver, Canada, Morgan Kaufmann, 1992:352- 362.

[3] Dell Zhang. "Semantic, Hierarchical, online clustering of Web search results"[C].Hangzhou China: Proceedings of The 6th Asia Pacific web conference, 2004:69-78

[4] Croft W B. "A model of cluster searching based on Classification". Information Systems, 1980, Vol.5:189-195

[5] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey" Advance Computing Conference, 2009. IACC 2009 IEEE International.

[6] L. Page, S. Brin, R. Motwani, and T. Wino grad, "The PageRank Citation Ranking: Bringing order to the Web". Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[7] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "Link Analysis: Hubs and Authorities on the World". Technical Report: 47847, 2001.

[8] Abou-Assaleh T., Das T., Weizheng G., Yingbo M., O'Brien P., Zhen Z., "A Link –Based Ranking Scheme For Focused Search".In:WWW2003, ACM Press.2007

[9] Bonnet 2001, Monica Bonnet, Personalization of web Services: Opportunities and challenges. http://www.ariadne.ac.uk/issue28/personalization/

[10] J.Jayanthi., K.S.Jayakumar., "An integrated Page Ranking Algorithm for Personalized Web Search". In International Journal of Computer Applications (0975-8887), Volume 12-No.11, January 2011.

[11] Callan, J.P., "Passage- level Evidence in Document Retrieval", In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994, Pp.302-310

[12] Yang, Y. and Zhang, H., "HTML Page Analysis Based On Visual Cues", In *6th International Conference on Document Analysis and Recognition (ICDAR 2001)*, Seattle, Washington, USA, 2001.

[13] D.Cai, S.Yu, J.-R.Wen, and W.-Y.Ma, "VIPS: a vision-Based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79, 2003.

[14] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype Document image analysis system for technical journals," Computer, vol. 7, no. 25, pp.10–22, 1992.

[15] K. Kise, A. Sato, and M. Iwata, "Segmentation of page I mages using the area Voronoi diagram," Computer Vision and Image Understanding, vol. 70, no. 3, pp. 370–382, June 1998.

[16] **"**Introduction to Information Retrieval" (HTML). Cambridge University Press. 2008. Retrieved 2008-11-09.