Survey on Different Ranking Algorithms Along With Their Approaches

Nirali Arora
Department of Computer Engineering PIIT, Mumbai University, India

Sharvari Govilkar
Department of Computer Engineering PIIT Mumbai University, India

ABSTRACT
Searching becomes a normal behavior of our life. Billions of users communicate with search engines daily. They are checking links of results, click on the ads, spend time on pages to restructure their queries and perform other actions. These interactions may concern some of the valuable source of information for tuning the content. There are massive web pages on the internet and search engines also have a test to find the best ranked pages. This paper provides a survey on different ranking algorithms such as link based ranking, content based ranking and usage based ranking and presents a comparative analysis on performance of these algorithms.

General Terms
Ranking, algorithms, query, similarity etc.

Keywords
Information retrieval, Content based ranking, Usage based ranking, Link based ranking, Web Mining, Page Rank Algorithms

1. INTRODUCTION
With billions of web pages available on the Internet. Search Engines always have a responsibility to find the best ranked order to satisfy user's query from those huge numbers of pages. A lot of search results that corresponding to a user's query are irrelevant to the user need. Considering the traditional IR scenario, a user structures a query and then triggers a retrieval process which results in a listing of ranked documents in decreasing order of relevance. An information retrieval (IR) model is a mathematical model that defines and explains major phases of the information retrieval process, including document text representation, user query representation, ranking of relevant documents. This paper is organized into 5 sections the section 2 provides the introduction to the ranking algorithms, section 3 discusses the major ranking algorithms the section 4 provides comparative study between the algorithms and then finally we conclude the paper in next section.

2. RANKING
The main focus of ranking algorithms is about using machine learning methods, such as classification clustering and regression methods, to proceed the task of ranking, to improve the ranking precision, recall and optimize the precision values of models. Learning and ranking are two processes to construct learning to rank models.

In the procedure of learning, given a large group of queries, ranking documents are retrieved as training data, and then a ranking model is built using the training data to obtain the finest model parameters.

After that, the ranking model can assign scores, which represents relevance and similarity between documents and queries, for testing documents and ranking them in a descending order of scores. At present most of the ranking based researches are based on the following approaches

2.1 Content Based Ranking
This approach is introduced for ranking the relevant pages based on the content and keywords rather than keyword and links provided by search engines. Based on the user query, search engine results are retrieved. Each and every result from the list of results are individually analyzed based on keywords and content. Preprocessing of user query is done to identify the root words. Each and every root words are considered for construction of dictionary and dictionary is built with set of synonyms for the user query. Every result page keywords and content words are pre-processed and compared against the dictionary. If a match is found, then accordingly weight is assigned to each word. Finally, the overall relevancy of the particular link against user query is computed by summarization of all the weights of the set keyword and content words.

2.2 Usage Based Ranking
Recommendation algorithms aims at providing "next" pages to the user based on her/his current visit and the past users' navigation and retrieval patterns. In the vast majority of ranking algorithms, only the usage data was used to produce recommendations, whereas the structural properties of the Web graphs were ignored. Usage-based ranking algorithms assign to score documents by how often they are viewed on Internet. For Usage-based ranking, there are limited works to provide the usage data in the web information retrieval systems, especially in the case of ranking algorithms. For some systems that do use the usage data in ranking, they determine the relevance of a web page by its selection frequency. This measurement is not that precise to indicate the real relevance. The time spent on reading the page, the task of saving, printing the page or adding the page to the bookmark, and the action of following the links in the page, are all good indicators, perhaps better than the simple selection frequency.

2.3 Link Based Ranking
To present the documents in a structured manner, web page ranking algorithms are applied which can arrange the documents in order of their relevance, importance and content score and use web mining techniques to order them. The link analysis algorithm is based on the linking structure of the documents. The quality of results from search engines is usually lower than what the user expects and this quality can be improved greatly if web pages are ranked according to some parameters based on links (in links) between the pages., i.e. a page which has many references and citations must have something to express. Link analysis based ranking algorithms are calculated offline and are usually static, before receiving any query from the user. This ranking often calculates popularity of a given web page by constructing web graphs using a set of nodes and analyzing the existing links in it. For example, the PageRank algorithm which is employed by

Google and Microsoft is one of the most important algorithms in ranking HTML web pages. The link analysis ranking that is generally calculated independent of query, enables search engines to have a faster and efficient retrieval to process a small amount of data. In fact, the results are sorted based on their query independent ranks and then top ranked list ones are selected accordingly. Web pages within a site do not exist in isolation they are usually related via hyperlinks. Link based ranking are divided into two types namely

1. Page ranking Algorithms

2. Hits Algorithms

2.3.1 Page Ranking Algorithms

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 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 has a high Page Rank. 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:

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

In eq.(1) the PageRank of page A is calculated where,
PR(A) = PageRank of Web Page A.
T1.....Tn =Web Pages that points to Page A (Inlinks of A)
C(A) = No of out link of webpage A
d = Damping factor its value is set between 0 and 1

Figure 1: Web graph used in page ranking algorithm [5]

2.3.2 Hits Algorithm

Jon Kleinberg introduced Hyperlink-Induced Topic Search (HITS)[6] (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 authoritative in the information that it holds, 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 represents a page that points to many other pages, and a good authority represented a page that is linked by many different hubs. 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.

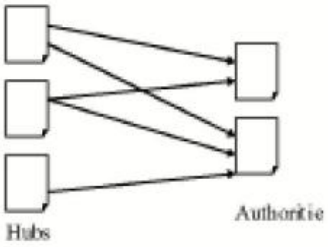

Figure 2 :Hubs And Authorities [6]

3. LITERATURE SURVEY

In this section we cite the relevant past literature that utilizes the various techniques for ranking. Ranking search results is a fundamental problem in information retrieval. Most common approaches focus on the similarity of query and page as well as overall page quality. However, with increasing popularity of search engines the capturing of user behaviors insists to appear on surface. A lot of methods have been done on implicit measures of user preference in field of information retrieval.

A novel usage based personalized page rank style algorithm was presented by Eirinaki et al [1]. This was used for ranking the web pages of the sites based on the user's previous navigational experiences and behaviour. Based on this algorithm a frequency selection for usage based ranking was developed. This algorithm can be mainly used to produce personalized recommendations.

The various problems in page ranking such as dangling links, historical values, false links Google Bombing, Google jacking etc were put forth by Gupta et al [2]. A concept of trust rank i.e. to place a core vote of trust on a set of seed set of reviewed pages was introduced. This trust rank can be bolted to produce more efficiency and high precision in obtaining the ranked pages

An efficient application of content based ranking on ranking blogs was suggested by Zhu et al [3]. In this approach the set of hyperlinks are divided into browsing links, recommendation links. A diagram of blog site is created by addition of hidden link based on the content analysis. A concept of implicit and explicit ranking was introduced in this technique.

A method proposed by Feyzania et al [4] to rank the documents by constructing a connected graph of semantic documents.This graph considers only the semantic links between the documents .Thus a novel approach of content based ranking was put forth accordingly by extracting the implicit ranks .Different weights can be used to distinguish the different semantic links.

An effective approach to usage based ranking using ontology was put forth by Junfang et al [5]. The weight calculation was done using ontology tree. This weight calculation takes into account of usage information, patterns and structure. The entire ontology tree was converted into a weighted graph and then justification was applied to calculate the final weight which was then used for searching, ranking and conflict solving

An important survey on page ranking algorithm was presented by Selvan et al[6] .The entire survey provided an efficient comparisons on these algorithms .The main focus was on the fundamental page ranking algorithms like hits and focused rank .The comparisons were drawn on the parameters of merits, demerits, performance and relevancy .

An effective approach of web service recommendation based on usage based ranking and QOS preferences was presented by Kang et al [7]. This algorithm is effective in classification and extraction of the results.

In [8] a novel approach in which crawler downloads only relevant documents taking advantage of migrants and thereby reducing the load on server it downloads only those web pages that are relevant to a particular topic or a set of topics and provides information in a personalized view considering only user preferences.

In [9] Patil et al has pointed out that the inference and analysis of search goals can have a lot of advantages in improving search engine relevance and user experience by combining web usage and web content mining.it presents a weighted technique to mine the web content catering to the user needs.

In [10] Kishorekumar et al proposed an efficient algorithm for mining the content through the extraction of search engine results and construction of dictionary for the user query and thus extraction of keywords that play a vital role in calculation of total strength of the keyword.

3.1 Inference Of Literature Survey

The following inference can be drawn on the basis of the literature survey.

- A link based analysis for semantic web documents is the best method for ranking regional language documents. This paper proposes a method to rank the documents by constructing a connected graph of the set of semantic documents.
- A link based page ranking algorithms used with timestamp is a temporal link analysis algorithm, it uses the timestamp of the nodes and links concerned, and it overall uses the weight of the pages, it reduces the number of the old pages and new pages raise in ranking result.
- Recommendation algorithms aims at proposing next pages to user based on current visit and past users navigational patterns. Usage based ranking algorithms present a novel technique of providing recommendations by analyzing the usage patterns.
- Despite the page rank algorithm is widely used today but it has various problems such dangling ranks, historical value, Google bombing, Google jacking, Google juice.
- Usage history can be used to determine the QOS of the patterns, user personalization rank combines the data and link analysis techniques for ranking and recommending web pages to end user.

4. COMPARATIVE ANALYSIS OF DIFFERENT RANKING ALGORITHMS

With the excessive growth of the information sources we are drowning in data but starving for knowledge ,therefore it has become necessary for the user to use information retrieval techniques and combination of different ranking algorithms to find and extract and filter the desired information.

Thus in order to broaden down the research on approach of ranking algorithms the following table provides the comparative analysis between the different ranking algorithms on various attributes and parameters.

Table 1 Comparison Of Different Ranking Algorithms

Sr no	Contrasting facts	Content Based Ranking	Usage Based Ranking	Link Based Ranking
1	Technique of mining	Web content mining	Web usage mining	Web structure mining.
2	Principle of ranking	Mining the content of the web pages	Discovering the useful navigation patterns	It discovers the link structure and the hyperlinks within the document
3	Process/concept	Uses the bag of words to represent the semi structured data	It extracts the useful information from the data in server logs, cookies maintained during the interaction with the user.	It is used to generate the structural summary in form of a web graph where web pages acts as nodes and hyperlinks as edges .

4	Ranking approach	Computes the rank value using the weighted approach to preprocess and rank the keyword present in the query.	Scores the document of how they are viewed by internet users .a lot of parameters are involved	Link based ranking computes the score on the basis of the web graph and degree of in links and out links.
5	Ranking process	It finds the similarity between the content and query , a cosine similarity is computed.	Usage based ranking is based on the time spent and the frequency of times the user visit the particular web page .	This provides a probabilistic distribution and it represents a likelihood that a person randomly clicking on the links will arrive at any particular page.
6	Performance	This provides a higher precision and recall the ' F' measure is high .	Medium performance and ranking as it is used for personalization	Less precision and recall as this algorithm ranks at the indexing time.
7	Quality of result obtained	Medium	High in only certain applications as usage characteristics are considered .	Performance varies from high to low depending on different applications
8	Input parameters	Inbound links, outbound links and content	Web server log, log files,ipaddress,cookies and session history	Inbound and outbound links
9	Complexity	<O(log N)	O log N	O(log N)
10	Search Engine	Used in IBM Search Engine to bring the best result	Used in web personalization and web recommender system .	Used especially in Google, AltaVista
11	Advantages	• Considers the priority to semantics • Suitable for tag based ranking	• Improves the user model without effort • Provides frequency based selection	• Highly efficient in finding a search of pages . • Demonstrates that it is good in finding the set of nodes
12	Disadvantage	• A suitable tuning is required	• Highly dependent on user behaviors ,modification and fraudulent behavior is possible	• Dangling links, Spider trap ,dead ends etc are possible . • Topic drift and irrelevant hubs are also possible.
13	Support to regional languages	Content based with semantic links are suitable for processing the regional languages.	No support in terms of regional language ranking	Lower support to regional languages ,support only in case of hyperlinks

5. CONCLUSION

Ranking is a crucial area of information retrieval domain. Developing an efficient searching engine in any domain requires designing an effective ranking technique which includes an efficient selection of parameters and characteristics. Usage based, Content based, Link based approaches for efficient ranking were explored. All the systems can retrieve relevant results from document. But to be specific, one cannot rank the different ranking algorithms in terms of performance as different approaches of ranking algorithms are suitable for different applications It is concluded that .Content based ranking algorithm is efficient for extraction of semantically rich content in regional languages, ontology ranking, tag annotation ranking .Usage based ranking is suitable for web personalization and recommender based systems .Link based is highly efficient for documents that have extensive hyperlinks.

6. ACKNOWLEDGEMENTS

A very special thanks to the computer department of Pillai college of Engineering Panvel for giving us the opportunity to conduct the research. This research survey wouldn't have been possible without the efforts of our principal Dr. RIK Moorthy .

7. REFERENCES

[1] Magdalini Eirinaki,Michalis Vazirgiannis Usage-based page rank for web personalizationin proceedings of the fifth IEEE international conference on Data mining (ICDM '05)

[2] Samriti Gupta,Alka Jindal Contrast of Link based Web Ranking Techniques at IEEE 2008

[3] Azam Feyznia,mohsin Kahanti A link analysis based ranking method for semantic web documents at ieee proceedings of 2010

[4] Jun fang,Lei Guo, Calculation of weight of entities in ontologies by using usage based information in ieee proceedings of *2011*

[5] Mercy paul Selvan,ChandraSekar ,A.Priya Darshan Survey on web page ranking algorithms in ijca (International Journal of Computer Applications) proceedings of 2012

[6] Guosheng Kang,JianxunLiu Active Web Service Recommendation Based on Usage History in ieee proceedings of service computing 2012

[7] Ashlesha Gupta,Ashutosh Dixit,AK Sharma Relevant document crawling with usage pattern and domain profile based page *ranking* in ieee proceedings of 2013

[8] P.Sudhakar,G.Poonkuzhal R.KishoreKumar A content based ranking for search engines in the IAENG proceedings of 2013

[9] Shital C Patil,RR keole Content and usage based ranking for enhancing search engine delivery *2014* in volume 3 issue 7International journal of Science and Research (IJSR)

[10] Safaa I. Hajeer, Rasha M. Ismail, Nagwa L. Badr, M. F. *Taiba* An Efficient Hybrid Usage-Based Ranking Model for Information Retrieval Systems and Web Search Engines . in ieee proceedings of 2015 6th International Conference on Information and Communication Systems (ICICS)

[11] "Metasearch engines and Information retrieval :Computational Complexity of ranking multiple search results" at fifth International Conference on Information Technology: New Generation 2008.1

[12] "Introduction to Information Retreival" (Html) Cambridge University press 2008

[13] "A Query Dependent Approach to Learning to Rank for Information Retrieval" for the Ninth International Conference on Web Age Information Management 2008.35

[14] Mandar kale,P Santhi Thelangam on Efficient Calculation of Page Rank in International Conference on Computer Science and Information Technology 2008

[15] Xiangi ,Chen Ding on QOS Based Ranking For Web Search in International Conference on Web Intelligence and Intelligent Agent technology 2008.