International Conference on Systems, Science, Control, Communication, Engineering and Technology 2016 [ICSSCCET 2016]

ISBN	978-81-929866-6-1	**VOL**	02
Website	icssccet.org	**eMail**	icssccet@asdf.res.in
Received	25 – February – 2016	**Accepted**	10 - March – 2016
Article ID	ICSSCCET020	**eAID**	ICSSCCET.2016.020

A Review on Ranking Algorithms in Information Retrieval

V Renupriya[1], G Saranya[2] and V Sakthipradeepa[3]

[1]Assistant Professor, Computer Science and Engineering, Karpagam Institute of Technology, Coimbatore.

[2]PG scholar, Computer Science and Engineering, Adhiyamaan college of Engineering, Hosur.

[3] Assistant Professor, Master of Computer Applications, Sri Vasavi College (SFW), Erode.

Abstract: Data mining can extract the useful and important information from a large database. The other important techniques in data mining related to extracting useful information are Text mining, Pattern mining, Information extraction, Information Retrieval etc. It is very difficult for a user to find high quality of document when there are many documents related to his search. User only wants the documents to be listed related to his query. Many ranking algorithms such as HITS, Weighted Page Rank, Dynamic Page Rank, Distance Rank and semi supervised ranking algorithm are used to display the ranked documents according to users need. In this paper the above listed algorithms are discussed and compared based on their ranking performance in information retrieval.

1. INTRODUCTION

The world is awash with information. Data mining refers to extraction of data from the large amount of information in database. Information retrieval is a popularized method related to data mining. Search engines are essential tools for finding the required information on the Web and other directories. The quality of a search engine is determined by the ranking functions which are used to provide the results according to users query. Ranking is the core of many information retrieval systems. The basic hypotheses behind Information Retrieval models are used for ranking the relevant and related documents according to user's query. Ranking algorithm [6] can overcome the problem of irrelevant document selection and gives the user the most informative document instead of displaying the entire document list.

The simple architecture of information retrieval system is shown in the Figure 1. There are three important processes in information retrieval they are, the Crawler also known as Robots that automatically download the web pages. The Indexer which is based on keywords. The indexer will match the keywords based on the user's query and retrieve the document. Finally before display the documents to the user the ranking mechanism is done to show the user the most informative documents with rank. The accuracy of ranking and its performance can be evaluated by Mean Average precision [7] and Discounted Cumulative gain [8].

Cite this article as: V Renupriya, G Saranya, V Sakthipradeepa. "A Review on Ranking Algorithms in Information Retrieval". *International Conference on Systems, Science, Control, Communication, Engineering and Technology 2016*: 102-105. Print.

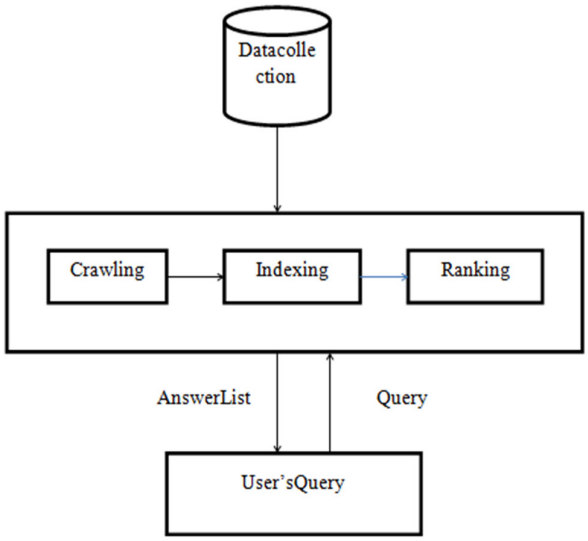

Fig 1. Information Retrieval System

2. Related Work

A wide variety of ranking algorithm for ranking document based on query has been proposed in the literature.Most existing algorithm are based on Pairwise, List wise[3]and Point wise approach which attempts to rank the documents based on comparison with other document. Martin Szummer [1] proposed a semi-supervised learning to rank algorithm. It learns to rank from both labeled data and unlabeled data.Extracting images from the web which uses SVM (Support Vector Machine) and Naive Bayes classifier algorithm which is proposed by Syed Hussain[4].The main idea behind the method is combining text or visual characteristicsfor automatic ranking of images according to query image. AdaRank[9] an boosting algorithm which is proposed by Xu and Li. Boosting is a general technique for improving the ranking performance and also it offers many advantages like easy implementation, efficiency in retrieval of relevant document and also accuracy in ranking.

3. Techniques

In this review paper the different ranking algorithm is discussed such as Dynamic Page Rank, HITS Rank algorithm, Distance Rank, Semi Supervised Rank, Weighted Page Rank

A. Dynamic Page Rank

KhaledAlsabti and Sanjay Ranka [5] proposed an algorithm Dynamic Page Rank, which identifies all the query words in the document. The query word can be enhanced by tokenization, Stemming, stop words removal as well as sense disambiguation approach. The resulted queries are passed to the search engine where the documents are retrieved based on the enhanced query.The web pages are ranked from higher to lower dynamic page rank values. The user receives more meaningful contents at top of the search results where least preferred documents are displayed at the last position in rank list.

B. Weighted Page Rank

NeelamTyagi [2] have analyzed that the World Wide Web, he proposed an algorithm Weighted page Rank(WPR) which is based on Visits of Links (VOL) being devised for search engines. The original WPR is the extension of standard Page Ranking algorithm[11]. This algorithm is proposed to display the web pages based on maximum visits that is the web page which is visited by many user will be displayed on the top position in rank list. So each and every page is given a page weight according the visits of the user.

C. Semi Supervised Learning Algorithm

Semi supervised learning algorithm which learns to rank from partially labeled data. In modern ranking system which uses fully labeled data where the quality of ranking is affected by the labeled data where it is time consuming and cost is high. The supervised learning system need labeled data and un labeled data where it contain large amount of unlabeled data than compared with labeled data. The Semi Supervised learning uses Lambda Rank [1]which directly optimizes popular retrieval metrics and improves retrieval accuracy.

Cite this article as: V Renupriya, G Saranya, V Sakthipradeepa. "A Review on Ranking Algorithms in Information Retrieval". *International Conference on Systems, Science, Control, Communication, Engineering and Technology 2016*: 102-105. Print.

D. Distance Rank

Ali Mohammad ZarehBidoki and Nasser Yazdani proposed a novel Distance Rank algorithm [10] based on recursive method. The algorithm is based on the distance factor where the distance between the web pages are calculated to compute the rank in search engine. The main advantage of this algorithm is it can find the pages faster and more quickly based on the distances solution. The distance rank algorithm adopts some properties of page rank. The page rank will have high rank value if it have more incoming link on a page.

E. Hits Algorithm

In 1999 proposed Kleinberg HITS(Hyper-Link Induced Topic Selection) algorithm which is query independent. In HITS ranking algorithm the web pages basically comes with two identities Hub and authority. The Hub which is used to collect all the web pages and acts as a pointer which points to various hyper-links. Authority which is used for the quality of the web pages and it is pointed by many hyper-links.

The algorithm concentrates only on the structure of the web not on their textual contents against a given query. The inlinks and outlinks of the web pages are considered during ranking.

4. Comparison of Ranking Algorithm

Based on the analysis on ranking algorithms the comparison is done based on the working process, limitation, search engine using algorithms and its drawbacks.

Table 1. Comparison between Ranking Algorithms

Algorithms	Working process	Search engine	Limitation	Drawback
Dynamic Page Rank	Computes the value during indexing and results are sorted based on priority of page.	Used in Google search engine.	The algorithm is query independent.	Even it is a good page if it don't have many link is considered as last position in rank list.
Weighted Page Rank	The result are sorted based on page importance.	Research model search engines uses WPR.	Query independent.	Provides high value ofrank to the more popular pages and it does not equally divide the rank of a page.
Distance Rank	The Average distance between two pages are calculated and rank the web page with minimum distance	Used in Research model	Efficiency problem	It has some Page Rank algorithm which affects the ranking process
Semi Supervised Learning	It is based on both Labeled and unlabeled data	Google and IBM search engine	Insufficient query selection	It tend to include non-informative documents when there are a large number of documents associated with each query
HITS	Highly relevant pages are computed first and sorted in rank list	IBM Search Engine	Blog ranking	The main drawback in HITS is Topic drift

5. Conclusion

Modern Search Engine such as Google, Yahoo, Bing and Ask etc., uses Ranking algorithm to retrieve useful document and web pages based on user's requirement. The users are interested only in top ranked documents instead of random display of documents which contains irrelevant information. Based on the literature study the Dynamic Page Rank and Weighted Page Rank shares the common ranking functionality of basic Page Rank algorithm. Google Search engine uses Page Rank algorithm which is efficient and low cost. HITS and Distance Rank is not so popular because of its efficiency problem and only its extensions are used in web pages.

Cite this article as: V Renupriya, G Saranya, V Sakthipradeepa. "A Review on Ranking Algorithms in Information Retrieval". *International Conference on Systems, Science, Control, Communication, Engineering and Technology 2016*: 102-105. Print.

Reference

1. Martin Szummer, "Semi-supervised learning to Rank with Preference Regularization", in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, pp. 662–663, 2011.
2. Wenpu Xing and Ghorbani Ali, "Weighted Page Rank ", Proceedings of the IEEE International Conference on Computer Science in 2004.
3. Xia F. Liu T. Wang J. Zhang W. and Li H. , "Listwise approach to learning to rank: theory and algorithm", Proc. 25th Int. Conf. Mach. Learn., pp. 1192–1199 2008.
4. Syed hussain S. and Kanya B.N., "Extracting Images from the Web using Data Mining Technique", International Journal of Advanced Technology and Engineering Research 2012.
5. KhaledAlsabti Sanjay Rankaand VineetSingh.,"Efficient Information Retrieval Using Dynamic Page Rank Algorithm", In Proceedings of IPPS/SPDP Workshop on High Performance Data Mining in 2011.
6. LaxmiChoudhary and Bhawani Shankar Burdak., "Role of Ranking Algorithms for Information Retrieval", Proceedings of the IEEE International Conference on Advance Computing. 2009
7. R. Baeza-Yates and B. Ribeiro-Neto. "Modern Information Retrieval". Addison Wesley, May 1999.
8. K. Jarvelin and J. Kekalainen. "IR evaluation methods for retrieving highly relevant documents". In SIGIR 23, pages 41–48, 2000
9. Xu J. and Li H. , 'AdaRank: A boosting algorithm for information retrieval', SIGIR, pp. 391–398, 2007
10. A. M. ZarehBidoki and N. Yazdani, "DistanceRank:An intelligent ranking algorithm for webpages" information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.
11. Rekha Jain, DrG.N.Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer application, Vol 13, Jan 2011.
12. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, no. 5, pp. 604– 632, 1999

Cite this article as: V Renupriya, G Saranya, V Sakthipradeepa. "A Review on Ranking Algorithms in Information Retrieval". *International Conference on Systems, Science, Control, Communication, Engineering and Technology 2016*: 102-105. Print.