A Complete Survey on Web Document Ranking

Shashank Gugnani
BITS-Pilani, K.K. Birla Goa Campus
Goa, India - 403726

Tushar Bihany
BITS-Pilani, K.K. Birla Goa Campus
Goa, India - 403726

Rajendra Kumar Roul
BITS-Pilani, K.K. Birla Goa Campus
Goa, India - 403726

ABSTRACT

Today, web plays a critical role in human life and also simplifies the same to a great extent. However, due to the towering increase in the number of web pages, the challenge of providing quality and relevant information to the users also needs to be addressed. Thus, search engines need to implement such algorithms which spans the pages as per user's interest and satisfaction and rank them accordingly. The concept of web mining tremendously assists in the mentioned scenario. Web mining helps in retrieving potentially useful information and patterns from web. This paper includes different Page Ranking algorithms and compares those algorithms used for Information Retrieval. Additionally it also presents some interesting facts about research in page ranking to find further scope of research in this area.

General Terms:

web document ranking, page rank

Keywords:

web structure mining, web content mining, web usage mining, document ranking

1. INTRODUCTION

With the size of the World Wide Web increasing at an exponential rate, it is becoming increasingly difficult to find relevant information. This main task of a search engine is to reduce this difficulty. It is the duty of a search engine to provide relevant information to the user on receiving a query. However, considering the size of the World Wide Web, a typical query might give more than a million results. The user does not have the time or patience to go through this huge list. Thus, ranking of web documents becomes a critical component of a search engine. Search Engines constantly need to find better and more efficient ranking methods, which can return high quality information to the user in as small a time frame as possible.

Search engines first create an index of all the web documents and store it on the server. After the user submits a query, the query is given to the index, which returns the documents containing the words in the query. Then, the returned documents are sent to a ranking function which gives a rank to each document and the top-k documents are returned to the user. Figure 1 shows the working of a typical search engine.

Web Mining is the task of extracting useful information from web documents. Web Mining comprises of three types: Web Structure Mining (WSM), Web Content Mining (WCM), and Web Usage Mining (WUM). Web Structure Mining uses the structure of the web, i.e. the hyperlinks between the web pages, Web Content Mining uses the content of the web documents and Web Usage Mining uses user click through data available in server logs. Every ranking algorithm employs a combination of one or more of these three types of Web Mining.

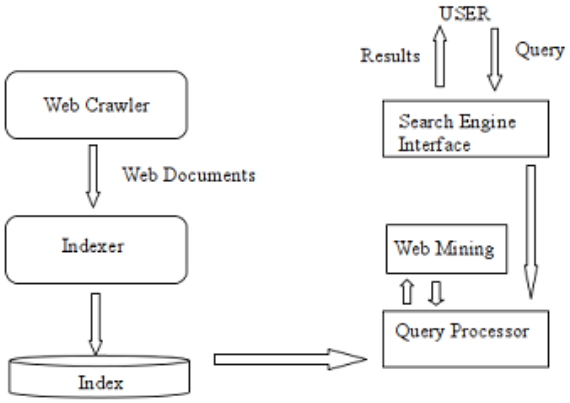

Fig. 1: Working of a Search Engine

The purpose of this paper is to list the important page ranking algorithms developed till date and compare their strengths, weaknesses, run time and efficiency so as to help in further research in this field. In addition, the page ranking algorithms have been compared according to 3 evaluation measures. Also, we have presented a summary of research work in ranking over the years and done an analysis of the same.

The rest of this paper is organized as follows. Section 2 enlists a summary of ranking algorithms arranged in ascending order of year to trace the development of ranking algorithms. Section 3 compares various ranking algorithms on a number of factors such as methodology, type of web mining, quality of results, etc. Section 4 compares the algorithms on the quality of their results based on three evaluation measures ($NDCG$, $P@n$ and MAP). Section 5 presents some interesting facts about research work in web document ranking and finally, Section 6 concludes the paper.

2. SUMMARY OF VARIOUS RANKING ALGORITHMS

Author/Year	Technique	Advantages	Limitations
Brin and Page, 1998 [5]	Graph based algorithm based on link structure of web pages. Consider the back links in the rank calculations.	More backlinks, higher the rank of a web page. Hence, authoritative pages are given preference.	Ranks are computed at indexing time not at the query time.
Kleinberg, 1999 [15]	Rank is calculated by computing hub and authorities score of pages in order of their relevance.	Returned pages have high relevancy and importance or link to pages with high relevancy.	Efficiency is less and there is a problem of topic drift.
Kim and Lee, 2002 [14]	This algorithm probabilistically estimates that clear semantics and the identified authoritative documents correspond better to human intuition.	Well defined semantics with clear interpretation. Efficiently provide answer to quantitative bibliometric questions.	Priority should be decided on the number of factors to model. Trades computational expense for the risk of only finding a local maxima.
Xing and Ghorbani, 2004 [22]	Based on the calculation of the weight of the page with the consideration of the outgoing links, incoming links and title tag of the page at the time of searching.	Higher accuracy in terms of ranking as it uses the content of the pages as well.	It is based only on the importance of the web page.
Baeza-Yates and Davis, 2004 [1]	This algorithm ranks the page by providing different weights based on relative position in page, tag where link is contained and length of anchor text.	It has less efficiency with reference to precision of the search engine.	Relative position was not so effective, indicating that the logical position does not always match the physical position.
Fujimura and Tanimoto, 2005 [8]	Use of the adjacency matrix, constructed from agent to object link not by page to page link. Three vectors, hub, authority and reputation are needed for score calculation of the blog.	Useful for ranking blogs because input and output links are not considered in the algorithm.	Only suited for blog ranking.
Bidoki and Yazdani, 2008 [4]	Based on reinforcement learning which consider the logarithmic distance between the pages.	Algorithm considers a real user by which pages can be found very quickly and with high quality.	If a new page inserted between the two pages, a large amount of computation needs to be done to calculate the distance vector.
Jiang et al., 2008 [9]	Visitor time is used for ranking. Sequential clicking is used for sequence vector calculation by using the random surfing model.	Useful when two pages have the same link structure but different contents.	Needs server logs for full efficiency.
Jie et al., 2008 [10]	The algorithm is based on the analysis of tag heat on social annotation web.	Ranking results are very exact and new information resources are indexed more efficiently.	Co-occurrence factor of tag is not considered which may influence the weight of the tag and decrease efficiency.
Keyhanipour et al., 2009 [13]	Creating a Genetic Programming Framework by using 13 Content based and Hyperlink derived features.	Low Computation time. More relevant results than standard algorithms.	Number of features used is less, which may lead to misclassification.
Lamberti et al., 2009 [16]	Ranks web pages for semantic search engine by using information extracted from the queries of the user and annotated resources.	Ranking task is less complex.	Every page is evaluated with respect to some ontology, which is computationally expensive.
Lee et al., 2009 [17]	Individual models are generated from training queries. A new query ranked according to the combined weighted score of these models.	It gives results for user's query as well as results for similar queries.	Limited numbers of characteristics for calculating the similarity.
Vojnovic et al., 2009 [20]	Suggests the popular items for tagging. Three randomized algorithms: frequency proportional sampling, move-to-set and frequency move-to-set are used.	Tag popularity is increased because large number of tags are suggested by this method.	Does not consider alternative user choice model, rules for ranking and suggestive rules.

Bhamidipati and Pal, 2009 [2]	Uses the score fusion technique.	It is used when two pages have same rank.	Does not consider the case when score vector T is generated from specific distribution.
Lian and Chen, 2010 [19]	Retrieval of moving object in the uncertain databases. It uses the P-Rank (probabilistic ranked query) and J-Prank (probabilistic ranked query on join).	It is very fast because it uses a R-Tree.	Experimental results are very promising only with limited number of parameters like wall clock time and number of P-Rank candidates.
Kayed et al., 2010 [12]	Builds ontology concepts using KAON (KArlsruhe ONtology). Uses ontology concepts to measure relevance for documents retrieved by top search engines.	Average ranking error is less than several search engines. Relevancy of results is enhanced by re-ranking.	The algorithm runs at query time which increases the response time of the search.
Bidoki et al., 2010 [3]	Aggregation of a number of ranking algorithms using Goodness Factor calculated based on user click-through data.	It combines the best of all ranking algorithms to produce highly relevant results.	The computation time of the algorithm is high.
Li et al., 2011 [18]	Generic rank aggregation framework consisting of building a Win/Loss graph of Web pages according to a competition rule, then applying the random walk on the graph and sorting these Web pages by their ranks using a PageRank like rank mechanism.	The results returned by the algorithm are authoritative and highly relevant to the query.	The content of the pages is not considered. Also, the algorithm runs at query evaluation time.
Zhu and Mishne, 2012 [23]	ClickRank: Using user click data as one of the features for ranking using machine learning.	ClickRank has a significantly lower computational cost than PageRank or BrowseRank. ClickRank delivers highly competitive ranking results.	Performance depends upon the features which are selected apart from ClickRank, which may cause misclassification, if a proper feature set is not used.
Du and Hai, 2013 [7]	By analyzing a user's browsing pattern and hyperlinks, the extension similarity and intension similarity are determined. Then by constructing an ISA and Part-Of hierarchy, the information content similarity between two nouns is computed automatically by using a user's web log.	It uses a combination of all 3 types of Web Mining. Hence, the results returned by the algorithm are authoritative and highly relevant to the query.	The algorithm runs at query time, which increases the response time of the search.
Wang et al., 2013 [21]	Text hypergraph for summarization and hypergraph based semi-supervised learning algorithm for sentence ranking.	Text hypergraph can integrate more group relations among multiple sentences. Only documents relevant to the query are retrieved.	Results are computed at query time, which increases the response time of the search. Also, importance of individual pages is not considered.
Jung and Lee, 2013 [11]	Multi-support vector domain description (multi-SVDD) to construct an efficient posterior probability function, for learning ranking functions.	Efficient utilization of memory and faster ranking. Ranking SVD outperforms ranking SVM in terms of practical ranking performance.	Importance of individual pages is not considered.
Derhami et al., 2013 [6]	In RL Rank, each web page is considered as a state and value function of state is used to determine the score of that state (page). A new hybrid approach using combination of BM25 as a content-based algorithm and RL Rank is used to rank documents.	RL Rank has higher performance in dense web graphs. RL Rank can achieve much better results than PageRank in standard criteria. The linear complexity of the RL Rank signifies the scalability of this algorithm on large datasets.	The algorithm runs at query time, which increases the response time of the search.

Table 1. : Summary of Ranking Algorithms

3. COMPARISON OF VARIOUS RANKING ALGORITHMS

Algorithm	**WPR**	**EigenRumor**	**Distance Rank**	**Page Content Rank**	**Topic Sensitive Page Rank**
Type of Web Mining	WSM	WCM	WSM	WCM	WSM
Methodology	Evaluates the values at indexing time and results are displayed in the sorted order as per the page's importance.	Computing the adjacency matrix between agent and object link.	Computing the minimum average distance between two pages and so on.	Evaluates new scores of the top n pages.	Evaluation as per the importance of contents.
Input Parameter	Backlinks and Forward links	Object or Agent	Backlinks	Content	Inbound and outbound links + the contents
Relevancy	Less(but Higher than PR)	High for Blogs	Moderate	More	More
Search Engine	Google	Research Model	Research Model	Google	Google
Quality of Results	Higher than PR	Higher than PR and HITS	Less than PR	Higher than PR	Much Higher
Advantages	Takes into account the importance of both the inlinks and outlinks.	Uses contribution of each community participant as well as each information object provided to the community.	Effect of the problem "rich get richer " is less as compared to PR.	Represents pages according to their content scores, unlike PageRank and HITS.	By calculating scores for different topics separately, the relevance of results is increased.
Disadvantages	Relevancy is compromised.	Used mostly for Blog Ranking, not for web page ranking.	Addition of new page leads to large calculation to calculate distance.	References are not considered.	Only applies with pages with texts in it (not for images and other attributes).

Table 2. : Comparison of Ranking Algorithms (I)

Algorithm	**Dirichlet Rank**	**Weighted Content Page Rank**	**Citation Count**	**Popularity Weighted Page Rank**	**Hypergraph-based Semi-Supervised Ranking**
Type of Web Mining	WSM	WSM&WCM	WSM	WSM	WCM
Methodology	Almost same as PR but uses Bayesian estimation for calculating probabilities.	Assigns weights to web links on the basis of relative position, tag and length of anchor tag.	Based on number of incoming citations.	Results are sorted according to weighted citations as well as popularity factor.	Incorporates the text hypergraph into the semi-supervised sentence ranking framework.
Input Parameter	Backlinks	Inbound and outbound links + the contents	Backlinks	Backlinks, Publishing time of paper	Document Set, Query
Relevancy	High	Less	Less	More(Less than PageRank)	High
Search Engine	Research Model	Google	Research Model	Research Model	Research Model
Quality of Results	High	Higher than PR	Less	Less	Higher than HITS
Advantages	Removes zero-one-gap problem present in Page Rank.	Both query relevance and page relevance is considered.	Simplicity of computation.	Popularity is taken into consideration.	Documents with high relevancy to query is given a high rank score.
Disadvantages	Works as a supplement for PageRank.	Judgment based on relative position is ineffective.	Unweighted Ranking.	Quality of publication is sacrificed.	The importance of individual pages is not considered.

Table 3. : Comparison of Ranking Algorithms (II)

Algorithm	**Multi-Support Vector Domain Description Probabilistic Generative Ranking**	**Formal Concept Analysis Semantic Ranking**	**A3CRank**	**Ontology Based Ranking**
Type of Web Mining	WSM & WCM	WSM, WCM & WUM	WSM, WCM & WUM	WCM
Methodology	The method utilizes multi-support vector domain description and constructs pseudo conditional probabilities for data pairs.	An extension similarity and an intension similarity that analyse a users browsing pattern. Semantic similarity between two concepts in two different concept lattices and finding the semantic ranking of web pages.	Aggregation of TF-IDF, DFR_BM25 and PageRank using OWA operator.	Measure the closeness (relevancy) of retrieved web sites to user query concepts and re-rank them accordingly.
Input Parameter	Web Documents	Initial query results, User's web log	Backlinks, Forward links, Document set, User's click-through data.	Initial query results, ontology domain.
Relevancy	High	High	High	High
Search Engine	Research Model	Research Model	Research Model	Research Model
Quality of Results	Less	High	High	High
Advantages	Efficient utilization of memory and faster ranking.	It uses all 3 techniques of web mining.	It uses all 3 techniques of web mining.	The average ranking error is low.
Disadvantages	The performance of a method significantly degrades when a ranking algorithm fails to locate a relevant document in a high-ranking position.	The algorithm runs at query time, hence the query execution time increases.	Execution time is high.	Building and maintaining ontologies for each users query will be expensive and it may not be applicable.

Table 4. : Comparison of Ranking Algorithms (III)

Algorithm	**Generic Rank Aggregation**	**Click Rank**	**CRLBM(RL Rank + BM25)**
Type of Web Mining	WCM	WUM	WSM & WCM
Methodology	Build Win/Loss graph of Web pages according to a competition rule, and then apply the random walk mechanism on the graph. Sort these Web pages by their ranks using a PageRank-like rank mechanism.	Click-Rank uses user click through data as one of the features to rank documents using machine learning.	Aggregation of RL Rank and BM25.
Input Parameter	Document set	Web logs	Backlinks, Forward links & Document set
Relevancy	High	High	High
Search Engine	Research Model	Research Model	Research Model
Quality of Results	Less	High	High
Advantages	Effective in facilitating users locating relevant information.	Delivers highly competitive ranking results.	This algorithm is better than basic algorithms (BM25, PageRank, and RL Rank) in quality of rankings.
Disadvantages	The importance of individual pages is not considered.	The importance of individual pages is not considered.	Since, it is an aggregation of two ranking techniques, computation time increases.

Table 5. : Comparison of Ranking Algorithms (IV)

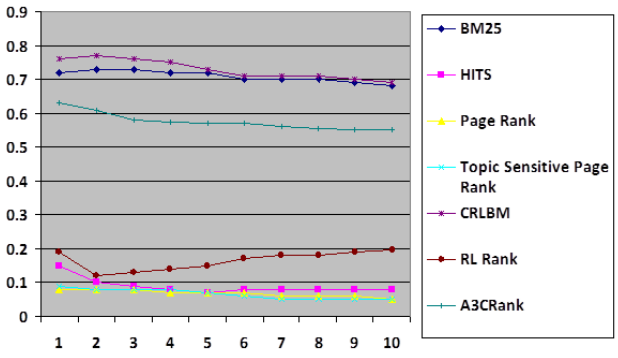

Fig. 2: P@n values of various ranking algorithms

4. EVALUATING QUALITY OF RESULTS OF VARIOUS RANKING ALGORITHMS

4.1 Evaluation Measures

Various methods are used for evaluation of query results. The most popular are Precision at n ($P@n$), Mean Average Precision (MAP) and Normalized Discount Cumulative Gain ($NDCG$). These three factors have been used to compare various ranking algorithms. There definitions are given below:
1. Precision at n ($P@n$): This is the ratio of top relevant documents to total number of documents (n) in the results.

$$P@n = \#\text{of relevant in top } n \text{ results}/n \tag{1}$$

2. Mean average precision (MAP): Average Precision (AP) corresponds to the average of $P@n$ values for all relevant documents of a given query.

$$AP = \sum_{i=1}^{n} (P@i.rel(i))/\#\text{total relevant docs for one query}, \tag{2}$$

where n is the number of retrieved documents, and rel(i) is a binary function on the relevance of the i^{th} document.
3. Normalized Discount Cumulative Gain ($NDCG$): Provides multiple levels of relevance Judgments. The $NDCG$ score of a ranking list for the first n positions with r_j as the rating of the j^{th} document in the ranking list is computed as

$$NDCG@n = \sum_{i=1}^{n} \frac{2r_j}{log(1+i)} \tag{3}$$

4.2 Comparison

Figure 2, 3, 4 give the $P@n$, $NDCG$, MAP values of various ranking algorithms respectively.

5. RESEARCH WORK IN WEB PAGE RANKING

The size of the Web is increasing at an exponential rate and relevant information is hard to find. Even with the increase in computational power and memory cost reduction, there is a need for better and faster algorithms for evaluating search results. Thus, research work in web page ranking is very important. From the graph in Figure 5 it is clear that the amount of research is increasing each year. This is a positive sign for the Information Technology Sector. By promoting research work in this field, finding relevant information on the web will never be a problem again.

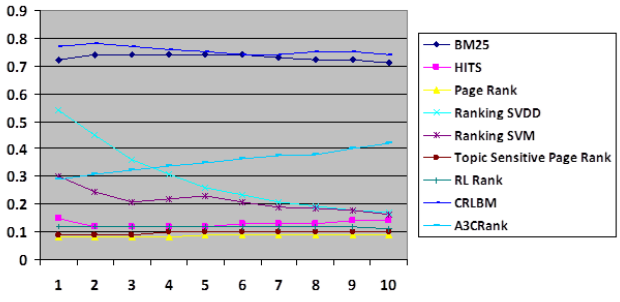

Fig. 3: NDCG values of various ranking algorithms

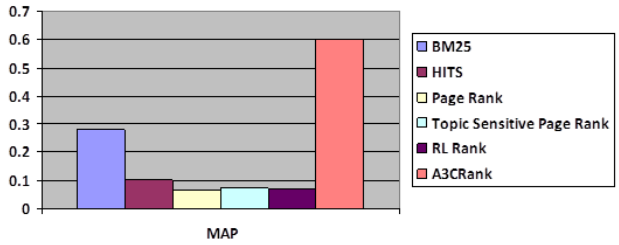

Fig. 4: MAP values of various ranking algorithms

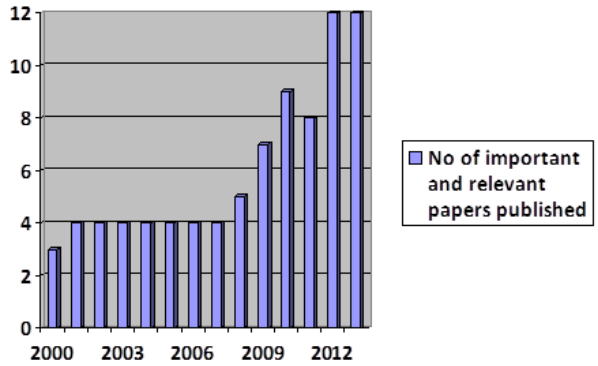

Fig. 5: Research work in Ranking

6. CONCLUSION

A large number of ranking algorithms have been proposed till date. This paper has summarized and compared some of these algorithms. Although each algorithm as its merits and demerits, we tried to find the best algorithm using various evaluation measures.

Based on the $P@n$ and $NDCG$ values, CRLBM (BM25 + RL Rank) was found to be the best ranking algorithm. According to MAP values, A3CRank was found to be the best algorithm.
With research work booming in this area, better algorithms are bound to emerge in the future and finding relevant data on the web will cease to be a source of concern.

7. REFERENCES

[1] R. Baeza-Yates and E. Davis. Web page ranking using link attributes. In *Proceedings of WWW-04and the 13th international World Wide Web conference - Alternate track papers & posters*, pages 328–329. ACM Press, 2004.

[2] Narayan L Bhamidipati and Sankar K Pal. Comparing scores intended for ranking. *Knowledge and Data Engineering, IEEE Transactions on*, 21(1):21–34, 2009.

[3] Ali Mohammad Zareh Bidoki, Pedram Ghodsnia, Nasser Yazdani, and Farhad Oroumchian. A3crank: An adaptive ranking method based on connectivity, content and click-through data. *Inf. Process. Manage.*, 46(2):159–169, 2010.

[4] Ali Mohammad Zareh Bidoki and Nasser Yazdani. Distancerank: An intelligent ranking algorithm for web pages. *Inf. Process. Manage.*, 44(2):877–892, 2008.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[6] Vali Derhami, Elahe Khodadadian, Mohammad Ghasemzadeh, and Ali Mohammad Zareh Bidoki. Applying reinforcement learning for web pages ranking algorithms. *Appl. Soft Comput.*, 13(4):1686–1692, 2013.

[7] Yajun Du and Yufeng Hai. Semantic ranking of web pages based on formal concept analysis. *Journal of Systems and Software*, 86(1):187–197, 2013.

[8] Ko Fujimura and Naoto Tanimoto. The eigenrumor algorithm for calculating contributions in cyberspace communities. In *Trusting Agents for Trusting Electronic Societies*, pages 59–74. Springer, 2005.

[9] Hua Jiang, Yong-Xing Ge, Dan Zuo, and Bing Han. Timerank: A method of improving ranking scores by visited time. In *Machine Learning and Cybernetics, 2008 International Conference on*, volume 3, pages 1654–1657. IEEE, 2008.

[10] Shen Jie, Chen Chen, Zhang Hui, Sun Rong-Shuang, Zhu Yan, and He Kun. Tagrank: A new rank algorithm for webpage based on social web. In *Computer Science and Information Technology, 2008. ICCSIT'08. International Conference on*, pages 254–258. IEEE, 2008.

[11] Kyu-Hwan Jung and Jaewook Lee. Probabilistic generative ranking method based on multi-support vector domain description. *Inf. Sci.*, 247:144–153, 2013.

[12] Ahmad Kayed, Eyas El-Qawasmeh, and Zakaryia Qawaqneh. Ranking web sites using domain ontology concepts. *Information & Management*, 47(7-8):350–355, 2010.

[13] Amir Hosein Keyhanipour, Maryam Piroozmand, and Kambiz Badie. A gp-adaptive web ranking discovery framework based on combinative content and context features. *Journal of Informetrics*, 3(1):78–89, 2009.

[14] Sung Jin Kim and Sang Ho Lee. An improved computation of the pagerank algorithm. In Fabio Crestani, Mark Girolami, and C. J. van Rijsbergen, editors, *ECIR*, volume 2291 of *Lecture Notes in Computer Science*, pages 73–85. Springer, 2002.

[15] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[16] Fabrizio Lamberti, Andrea Sanna, and Claudio Demartini. A relation-based page rank algorithm for semantic web search engines. *Knowledge and Data Engineering, IEEE Transactions on*, 21(1):123–136, 2009.

[17] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, and Shie-Jue Lee. A query-dependent ranking approach for search engines. In *Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on*, volume 1, pages 259–263. IEEE, 2009.

[18] Lin Li, Guandong Xu, Yanchun Zhang, and Masaru Kitsuregawa. Random walk based rank aggregation to improving web search. *Knowl.-Based Syst.*, 24(7):943–951, 2011.

[19] Xiang Lian and Lei Chen. Ranked query processing in uncertain databases. *Knowledge and Data Engineering, IEEE Transactions on*, 22(3):420–436, 2010.

[20] Milan Vojnovic, James Cruise, Dinan Gunawardena, and Peter Marbach. Ranking and suggesting popular items. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8):1133–1146, 2009.

[21] Wei Wang, Sujian Li, Jiwei Li, Wenjie Li, and Furu Wei. Exploring hypergraph-based semi-supervised ranking for query-oriented summarization. *Inf. Sci.*, 237:271–286, 2013.

[22] Wenpu Xing and Ali A. Ghorbani. Weighted pagerank algorithm. In *CNSR*, pages 305–314. IEEE Computer Society, 2004.

[23] Guangyu Zhu and Gilad Mishne. Clickrank: Learning session-context models to enrich web search ranking. *TWEB*, 6(1):1, 2012.