A New Technique for Ranking Web Pages and Adwords

K. P. Shyam
Computing Science and Engineering
VIT University – Chennai Campus
Vandalur Kelambakkam Road
Chennai, India

Sharath Jagannathan
Computing Science and Engineering
VIT University – Chennai Campus
Vandalur Kelambakkam Road
Chennai, India

Maheswari Rajavel, Ph.D
Computing Science and Engineering
VIT University – Chennai Campus
Vandalur Kelambakkam Road
Chennai, India

ABSTRACT
Web mining is an active research area which mainly deals with the application on data mining techniques on the data that is provided by the internet, the World Wide Web(WWW).The Information provided by the internet could be in either webpages, links structure of WWW or Web server logs. Web content mining, Web usage mining and Web structure mining are the three categories by which web mining is classified. This paper proposes a technique by which the search results can be refined in such a way that the results provided to the user are unique and the best suited result. This is achieved by using a new technique known as the Semantic Rank (SR) algorithm. The SR algorithm ranks the webpages in a more efficient way than the PageRank algorithm used by Google.

Key terms
Web Mining, Page rank, Search engine, Semantic rank.

1. INTRODUCTION
Search engines use ranking algorithms for fetching information from the Internet. A Search engines does not favor the user all the time. The search engine fetches no proper information always.

For example, user intentions may vary widely with different users. A user enters the keyword "A", he/she may want to know about the features of the keyword and another user may want to know the cost of the product. There are varied user intentions. Not all the search engine provides the accurate results. The user interaction with the World Wide Web(www) is shown in Figure 1.

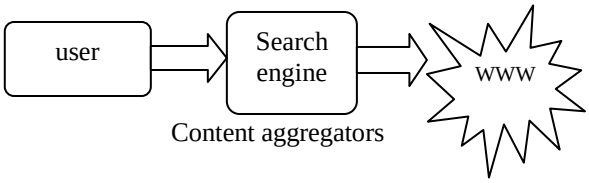

Figure 1.Concept of search engine

The Search engine does not provide the user with the appropriate results requested by the user always. For improving the search results, many ranking algorithms have been proposed.

The Query processor component in the search engine retrieves the related information. The URLs are fetched by the query component after indexing. But, the search engine performs the ranking mechanisms to make the results easier for the user before presenting.

In this paper a different approach for fetching the appropriate information according to the user's search intention is proposed. The web pages are traversed using several links in the web pages. Considering a web page and the user uses a link in that particular we page in order to move to the next page and also taking into account the number of out degree of links in the traversed pages. This is the Markov Decision Process (MDP) [8]. Therefore the reinforcement learning in this process includes the web pages (states), the out degree of links in each page (Action), the surfer clicks one of the links to navigate to the next page (Policy), the inverse of out degree of source page (Reward) and the total number of pages that the surfer traverses during the state (Value Function). Based on this the pages in a search engines are ranked rather the Page Rank method.

2. RELATED WORK
2.1 Browsing the web
Web browsing can be performed in many ways. A wide range of web mining tools can determine popular keywords. Zero-Click [1] and zooming cross media [2] allow users to browse the Web content without clicking. A preview of the link is displayed when the cursor is placed on it. But these techniques do not re-rank the search results.

2.2 A Typical Web Search
There are many ways to search in a web browser. But how to display the results back to the user must be taken in consideration. There are many studies to analyze the user's intentions based on their search in a search engine. Tian et al [3] used multiple keywords and analyzed them based on their semantic relationship and proposed an advanced web retrieval method by using proximity.

When the user enters the keyword in the search engine, two processes takes place. First is the front-end process. The query processor takes in the keyword. The Query processor fetches the information related to the keyword in the database. The Back-end process performs the ranking of the results. The resulting information is provided to the user. A simple architecture of a search engine is shown in Figure 2.

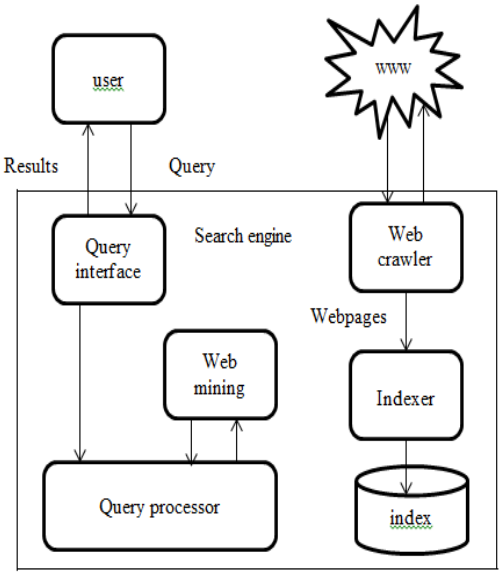

Figure 2: Architecture of search engine

2.3 Adwords in search engines

A Search engine provides the user with relevant results by matching the query given by the user with millions of results provided in the web. The Order in which the searcher gets the resulting pages for a query is based on the indexing of text on those pages, text in links pointing to those pages, and some measure of importance based upon link popularity.

The Adwords mainly depend on the keywords that the user enters in the search engine. The Quality score determines the position of the ad in a webpage. The Phrase keyword of the ad must match with the keyword that the user searches. The Adrank can be determined by calculating the quality score and the cost per click.

Quality score x Max cost per click=Adrank

The Score can be improved by using the keyword as the title. Adding information such as phone numbers, area and city can improve the score because there is always a higher probability that the user search according to their location.

The quality score of the page can be found in a cycle of steps as shown in the Figure 3.

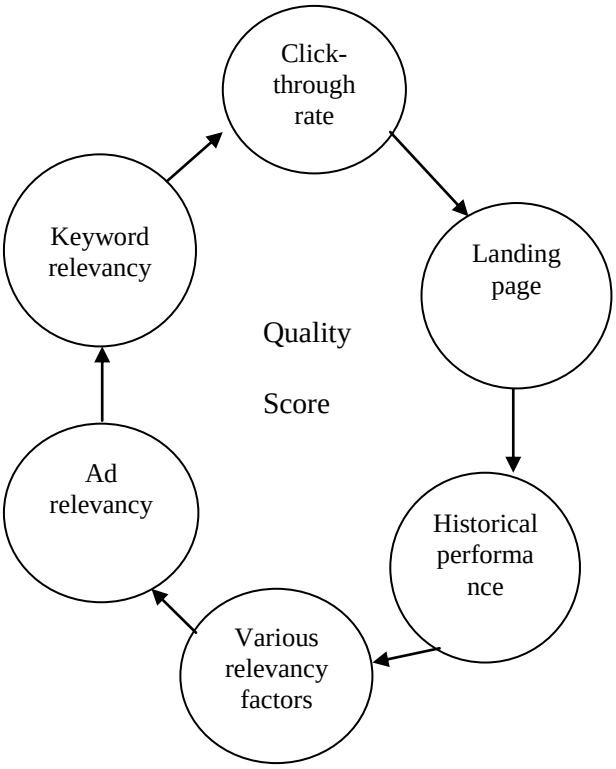

Figure 3: Quality score of adwords

3. RERANKING METHODS

3.1 Page Rank Algorithm

Sergey Brin and Larry Page[4,5] developed the PageRank algorithm. Google uses this algorithm. It was named after Larry Page(co-founder of goggle).that uses the link structure of the web to determine the importance of web Pages. Google's PageRank algorithm works in such a way that, a page is considered important when the incoming and the out going links to and from that page is high. Therefore, the PageRank algorithm takes the links in account and does the ranking operation on the links. Thus, a page obtains a high rank if the sum of the ranks back links is high.

PageRank algorithm takes around 25 billion web pages into account on the WWW to assign the rank [6]. Google has all the ranks precompiled and it compares with the text matching scores in order to obtain a ranking score for each resulting web page in response to the given query.

A simplified version [11] of PageRank is defined in Eq. 1:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (1)$$

Where u represents a web page, B(u) is the set of pages that point to u, PR(u) and PR(v) are rank scores of page u and v respectively, Nv denotes the number of outgoing links of v, c is a factor used for normalization.

In PageRank, the rank score of a page (say p) is equally divided based on the outgoing links from p. The Outgoing links of page

p are assigned values which are in turn used to calculate the ranks of the pages pointed to by p.

But not all the users use direct links on the WWW. So the PageRank was modified and is given by Eq. 2 :

$$PR(A) = (1- d) / N + d(PR(T1) / C(T1) + .. + PR(Tn) / C(Tn)) \quad (2)$$

Where, PR(A) is the PageRank of page A and PR(Ti) is the PageRank of pages Ti which link to page A.C(Ti) is the number of outbound links on the page Ti and d is the damping factor whose value is set between 0 and 1.

3.1.1 Example illustrating work of PR

The characteristic of PageRank is illustrated by a small example as shown in Figure 4.

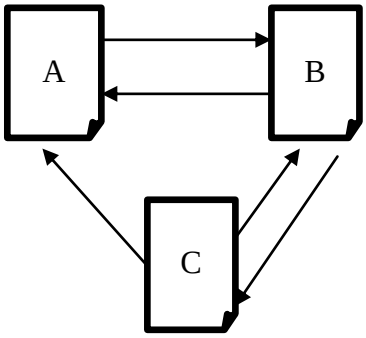

Figure 4.Example of Hyperlinked structure

Consider a small web consisting of three pages A, B and C, where page B links to the pages A and C, page A links to page B and page C links to page B and A. According to Page and Brin, 0.85 is the damping factor that is set usually, but we set it to 0.5 to make the calculations simple. By using the exact value of the damping factor, the PageRank algorithm has minor effects, but the fundamental principles of the PageRank are not influenced by it. So, we get the following equations for the PageRank calculation:

$$PR(A) = 0.5 + 0.5PR(B)$$
$$PR(B) = 0.5 + 0.5(PR(A) / 2 + PR(C))$$
$$PR(C) = 0.5 + 0.5(PR(A) / 2 + PR(B))$$

Solving the above equation we get the following PageRank values for the single pages:

PR(A)=14/13=1.07692308
PR(B)=10/13=0.76923077
PR(C)= 15/13 = 1.15384615

In this example, it is obvious that the PageRank can be given by the sum of all the pages and thus it equals the total number of web pages.

It is very easy to find the PageRank of a simple web consisting of three pages. But in practice, a web contains billions of web pages and it is not possible to find a page just by inspection.

3.2. The Proposed algorithm:

SR Rank algorithm inspired from reinforcement learning concepts. So in this section, we first review reinforcement learning concepts. Afterwards, SR Rank algorithm is introduced.

3.3 SR Rank algorithm

Using the reinforcement concepts and the link structure of the web pages, the pages are ranked. In SR Rank algorithm the Agent is the surfer of the wed and the State is the each web page. The surfer (agent) clicks on the any available link in each page (state) and traverse the web pages in an uniform probability and goes to the next state.

Therefore, the agent clicks randomly on the available links in order to traverse the web pages with a uniform probability. In other words, when an agent selects a link by clicking randomly on one of the available link in the current state, then the policy π is equal to 1/O , where O denoting the current state is the out degree of the current state. When a transition occurs from j to i , where j is the current state and i is the next state, a reward is given by Eq. 3:

$$r_{ij} = 1/O(j) \quad (3)$$

Where O(j) is the out-degree of page j. Reward is more to the page where the out degree from the page is less. The score of the page i can be defined by sum of the discounted rewards calculated by the agent during traversing the pages to reach page i. Then the accumulated rewards are discounted by the Agent r_{ij} . Therefore the score of the page i can be calculated by probability of reaching the state i multiplied by the sum of the rewards accumulated and the total transition rewards. The score of page can be calculated by Eq. 4:

$$R_{t+1}(i) = \sum_{j \in B(i)} (prob(j) / O(j)) \cdot (r_{ij} + \gamma R_t(j)) \quad (4)$$

Where, $R_{t+1}(i)$ - rank of page i in time t + 1
$R_t(j)$ - the rank page j in time t,
B(i) - set of pages that point to page i
prob(j)- probability of the agent at page j
O(j) - out-degree of page j
r_{ij} - reward for transition from page j to i.

Therefore, by Eq.(4), the rank of the page i can be calculated by the out degree and the rank of the pages pointing to the page i. By using the policy evaluation concept [8] in the reinforcement learning, the rank of a page can be estimated.

3.4 Best fit algorithm

By the Best fit algorithm, only webpages and the results that are ranked the highest are given to the user instead of the user's intentions. The resulting webpages may or may not be relevant to the user. In this way that Ads relating to the user's search can be provided in more efficient way. Every time when the user enters a new keyword it allocates the ads using the Best fit algorithm.

3.5 Apriori algorithm

Another method such as the Apriori algorithm can be used in order to refine the search. When the user enters the keyword, the Ads for the respective keyword are generated in such a way that it provides the user with the information of the respective keyword and also the next probable result that the user would

like to search. For example, If the user searches "pet medicine", Ads like pet clinics, medicines for pets are generated. The user is also provided with the Ads which are closely related to the keyword like pets adoption, dogs care.

In this way instead of the results of Ads that appear according to the auction or bid done by the advertiser, the results can be provided to the user based on the keyword and the user's search intention.

4. EXPERIMENTS

Comparison of the Page Rank algorithm and the SR Rank algorithm are evaluated by using sample datasets in order to find the efficiency of both the algorithms. Page Rank algorithm only provides the user with the results containing the highest rank whereas the SR algorithms provide the users with the results determining the links availability in a particular page.
The first experiment includes the Page Rank algorithm and the SR Rank algorithm as connectivity based algorithms. The SR Rank was set to γ=0.9 and the damping factor in the Page Rank is set to 0.85.

By using the best fit and the apriori algorithms, the adwords that are generated are more relevant to the user's search intention rather than the generation based on the advertiser's auction. In this way, the user can get the ad results relevant to their search.

4.1 Experimental results

The dotIR benchmark datasets are shown in Figs. 6–8. Figs. 6 and 7 show the obtained P@n and NDCG@n, respectively. As shown, the obtained values for SR Rank are higher than those for PageRank, especially P@1 and NDCG@1. Fig. 8 shows that SR Rank obtains improvement about 26% over the PageRank in terms of MAP measure.

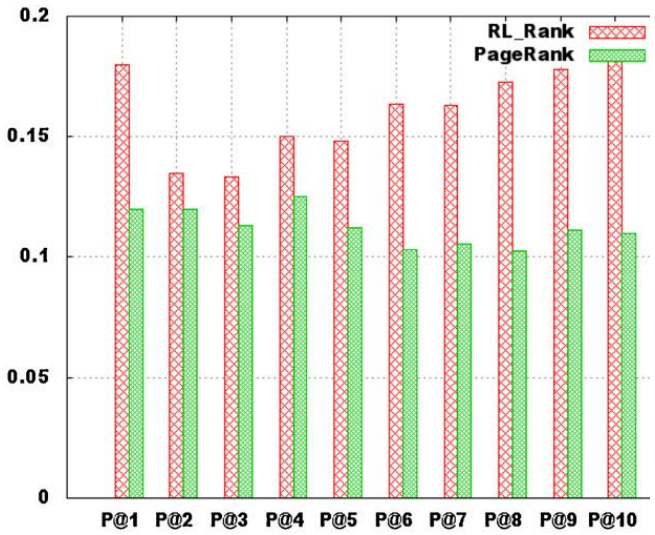

Figure 6: SR rank and PageRank in dorIR benchmark dataset

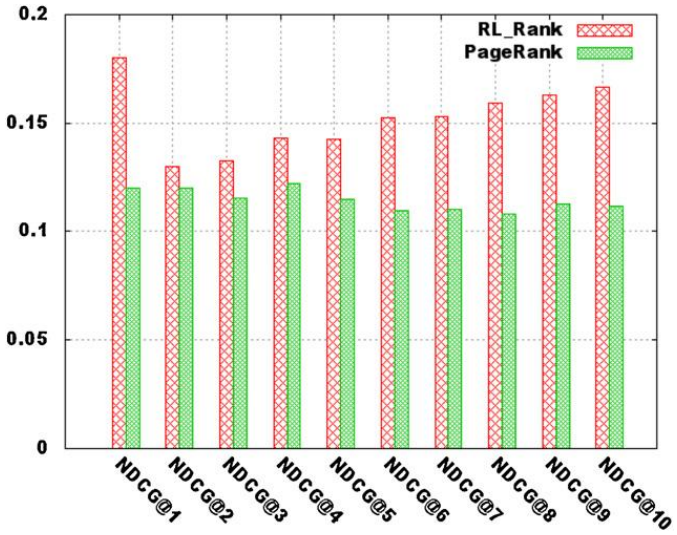

Figure 7: SR Rank with PageRank in the NDCG@n measure on dotIR benchmark.

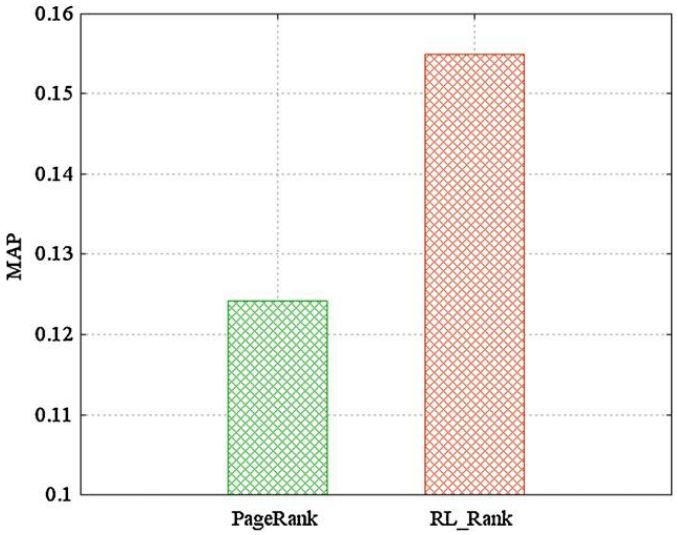

Figure 8: PageRank and SR Rank in MAP measure in dotIR benchmark

5. CONCLUSIONS AND FUTURE WORK

Comparison of the Page Rank algorithm and the SR Rank algorithm are evaluated by using sample datasets in order to find the efficiency of both the algorithms. Page Rank algorithm only provides the user with the results containing the highest rank whereas the SR algorithms provide the users with the results determining the links availability in a particular page. In future, the efficiency of the SR algorithm can be further improved and a few security measures can be added.

6. REFERENCE

[1] T. Nanno, S. Saito, and M. Okumura. Zero-Click: a system to support Web browsing. The 11th international conference on WoSRd Wide Web, 2002.

[2] T. Araki, H. Miyamori, M. Minakuchi, A. Kato, Z. Stejic, Y. Ogawa, and K. Tanaka. Zooming Cross-Media: A Zooming Description Language Coding LOD Control and Media Transition. Proceedings of the 16th International

Conference on Database and Expert Systems Applications, LNCS, Springer, pages 260–269, 2005.

[3] C. Tian, T. Tezuka, S. Oyama, K. Tajima, and K. Tanaka. Improving Web Retrieval Precision based on Semantic Relationships and Proximity of Query Keywords. Proceedings of the 17th International Conference on Database and Expert Systems Applications, LNCS, Springer, pages 54–63, 2006.

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

[5] http://www.google.com/technology/index.html,Our Search:Google:The Technology.

[6]http://WWW.webrankinfo.com/english/seo-news/topic-16388.htm. January 2006, Increased Google index size.

[7] Samarati.P, "Protecting Respondents Identities In Microdata Release, IEEE Trans.Knowl.DataEng.13 (6):1010-1027(2001).

[8] R.S. Sutton, A.G. Barto, Reinforcement Learning. An Introduction, MIT Press, Cambridge, MA, 1998.