Comparison of Link Based Web Page Ranking Algorithms based on Weighted Graph using Probabilistic Approach

RAVINDER KUMAR
ravinder@thapar.edu

PREET KAMAL
deol.preetkamal@gmail.com

Department of Computer Science & Engineering
Thapar University, Patiala – 147 004 (India)

Abstract. The web today plays an important role in the cultural, educational and commercial life of millions of the users. With the huge amount of information available on the web, users typically rely on the web search engines in order to get the most desired and relevant information. As most of the users examine the first few pages so the key for user satisfaction is to give the desired results in the first few pages. Therefore the role of ranking algorithms is crucial i.e. select the pages that are most likely be able to satisfy the user's needs and also bring those results to the top positions. These ranking algorithms use a web graph as an input having crisp values. When the refined values of the web graph are used the performance of the algorithm is improved. The refined values of the web graph are obtained by calculating the conditional probability of each out links. The performance measures for the ranking algorithms are Mean Reciprocal Rank, Mean Average Precision and Normalized Discounted Cumulative Gain values. The rank values are calculated and their efficiency is compared with the present scenario of considering the crisp values of the out links to the proposed scenario of considering the conditional probabilities of out links.

Keywords: Hyperlinks, ranking, Page rank algorithm, HITS algorithm, SALSA algorithm, MRR, MAP and NDCG value.

1. Introduction

The goal of information retrieval is to find out all the relevant documents for a user query in a collection of documents. With the advent of the web new sources of information became available [1]. One of them is the hyperlinks between documents and records of user behavior. To be precise, hypertexts (i.e., collections of documents connected by hyperlinks) have existed and have been studied for a long time [1]. What was new was the large number of hyperlinks created by independent individuals. This area of information retrieval is commonly called link analysis.

One of the fundamental problems in Information Retrieval is the ranking of search results. In the context of web search, where the corpus is massive and queries rarely contain more than three terms, most searches produce hundreds of results. After receiving a query from user, the URL's of the documents are returned by the search engines to the user in the decreasing order of relevance of these documents, hence this process is known as ranking. Given that the majority of search engine users examine only the first page of results [2], effective ranking algorithm is the key to satisfying users' needs.

Ranking of a webpage depends on many factors such as age of the webpage, loading time, HTML code error, hosting service, dynamic business world, quality and quantity of the content, links towards and from the webpage, domain name or URL of site registration. So keeping all these factors in mind and determining the rank of a webpage is a tedious task. In order to rank various webpages various ranking algorithms have been proposed.

Over the past decade, there has been an abundance of research on link-based ranking algorithms. Most of this research has centered on proposing new link-based ranking algorithms or improving the computational efficiency of existing ones (primarily PageRank [3]), but there are very few studies on validating the effectiveness (ranking performance) of well-known algorithms on real and large- scale data sets. We believe that this is primarily due to the fact that conducting such studies requires substantial resources: large web graphs, which are typically obtained by crawling a substantial portion of the web; query logs, which are hard to obtain from commercial search engines due to privacy concerns; and human relevance judgments of result sets, which are expensive to produce. It becomes important to develop the IR evaluation techniques and approaches which incorporate the ability of retrieving highly relevant results from the user perspective IR methods when there is a need to develop Information retrieval techniques in this direction.

Users are often interested in the most relevant results and ignore the other results. Thus it becomes necessary for the ranker to spend effort in getting the topmost results right. Various performance measures have been developed in the information retrieval field which is paying special attention in bringing the most relevant results at the topmost ranking. Examples of such measures are mean average precision, mean reciprocal rank, and normalized discounted cumulative gain.

This paper emphasis on determining the relative effectiveness of the ranking algorithms namely Page Rank, HITS[4], SALSA[5,6] and Norm (p)[7] algorithms on the basis of performance measures namely mean average precision[8], mean reciprocal rank[9], and normalized discounted cumulative gain[10]. It also exploits the feature of web graph where webgraph is represented as an adjacency matrix in the ranking algorithm. When a page A has an outlink to another page say B it is represented in the adjacency matrix as 1.In the adjacency matrix if there is a link between two webpages it is represented as 1 else 0. It does not consider the fact that webpage A may have other outlinks as well like those of C and D besides B. When a user is on page A he has the freedom to visit any of the page from B, C and D. So the chances of navigating to any of the page from A are equally divided among B, C and D. In this study we incorporate this feature and assign probability to the outlinks instead of assigning 1.

The paper is structures as follows: section 2 contains the study of the related work, section 3 gives an insight into the data set used to carry out the study, section 4 details the performance measures used, section 5 reviews the various ranking algorithms like those of Page Rank, HITS ,SALSA and Norm (P) algorithms, section 6 presents the experimental results and section 7 provides the conclusion of the study.

2. Related Work

Page Rank is a query independent algorithm, which is based on the connectivity structure of the web pages. It is used by Google search engine. The Page Rank value of a page is weighted by each hyperlink to the page proportionally to the quality of the page containing the hyperlinks; i.e., the Page Rank value of a page will spread evenly to all the pages it points to[3]. The Page Rank of a web page is therefore calculated as a sum of the Page Ranks of all pages linking to it (its incoming links), divided by the number of links on each of those pages (its outgoing links).

HITS (Hyperlink-Induced Topic Search) a query-based algorithm. From the user query, the HITS algorithm first creates a neighborhood graph for the query. The neighborhood graph contains top 200 matched web results retrieved from a content-based web search engine. It contains all the pages of the 200 web pages linked to and web pages that linked to these 200 top pages. After this an iterative calculation is performed on the values of authority and hub. For each webpage p, the authority and hub values are calculated. The authority value of webpage p is the sum of hub scores of all the webpages that points to p, the hub value of page p is the sum of authority scores of all the webpages that p points to . Iteration proceeded on the neighborhood graph until the values converged.

The SALSA algorithm combines the ideas both from page rank and HITS algorithm. In this algorithm, a random walk on the bipartite hubs and authorities graph alternatively between hubs and authorities is performed. When on an authority side of the bipartite graph at a node, the algorithm selects one of the incoming links uniformly at random and moves to a hub node on the hub side. When at node on the hub side the algorithm selects one of the outgoing links uniformly at random and moves to an authority.

Norm is a function which assigns a positive length to all the vectors in the given vector space. These algorithms belong to the family of additive online learning algorithms. These work on the principle of preferential treatment of the authority weights. It can be implemented by using a norm or an operator. By this we will be able to use the fact that lower authority weights contribute less to the hub weight. The simplest approach is to scale the weights. Now the question is how to choose the scaling factors. The most common solution to this question is to use the authority weight to determine the scaling factor. As higher authority weights are significant in the calculation of hub weight so hub weight of the given node i is set to be the p-norm of the vector of the authority weights of all the nodes pointed to by the given node *i*.

We are familiar with the two earlier studies that assess the performance of SALSA and compared it with those of HITS, Page Rank and in degree. This study is presented by Borodin et al. This study was carried on 34 queries having result set of 200 pages per query which was obtained from Google[11]. The neighborhood graph was derived by retrieving the backlinks. These backlinks suffers from some limitations like it returns some backlinks and by no means returns a uniform random sample. The second study was performed by Marc Najork on a set of over 28,000 queries and a webgraph containing around 3 billion URLs[12].

3. Data Set

For the purpose of this study a webgraph having 20,000 nodes which are interconnected have been selected. These 20,000 nodes are subdivided into 20 nodes which are named from {A to T} in order to manage and calculate their respective values effectively. On this webgraph PageRank, HITS, SALSA and Norm(P) algorithms are applied. When the rank of these algorithms are calculated the performance measures like those of MAP,MRR and NDCG values are determined. Determining the retrieval results from the webpages scores and human judgment is not very common and are the subject of research on the field of information retrieval. An efficient performance measure should take into account the user satisfaction keeping the fact that they will not like to dwell deep to obtain the desired results.

4. Measure of effectiveness

The present study has been conducted using the performance measures like those of MRR, MAP and NDCG. These performance measures compare Page Rank, HITS, SALSA and Norm (P) algorithms. These algorithms are compared on the basis of value of the adjacency matrix value $m_{i,j}$. For the given webpages Wi and Wj the adjacency matrix M = ($m_{i,j}$) defined as

$$m_{i,j} = \begin{cases} 1 & ifWi \rightarrow Wj \\ 0 & otherwise \end{cases}$$

In this study we have used the probability of visiting the webpage Wj from the webpage Wi in the adjacency matrix instead of 0 and 1.In the webpage Wi all the links are analysed and hence the probability of visiting the given hyperlink is calculated. If there are 4 hyperlinks on the webpage Wi then the adjacency matrix is defined as

$$m_{i,j} = \begin{cases} \frac{1}{numberofoutlinksonWi} ifWi \rightarrow Wj \\ 0 \quad otherwise \end{cases}$$

In this section we will briefly review various performance measures and their respective definitions of Mean average precision, mean reciprocal rank, and normalized discounted cumulativegain. Given a rank ordered rank vector of n results, let rat (i) be the rating of the result assuming 3 being the “definitive” and 0 being “detrimental”

4.1 Mean Average Precision

It is one of the most frequently used measures of a ranked retrieval run. To define MAP one needs to define Precision at position k (P@k). The Average precision of a given query is the arithmetic mean of the precision scores after each relevant document is retrieved. It takes into account both recall and precision oriented aspects. The precision P@k at document cut off value k is defined the fraction of relevant results among the k highest ranking results and is represented as

$$\text{Precision P@k}=\frac{1}{k}\sum_{i=1}^{k} rel(i)$$

The average precision at document cut off value of k[8] is defined to be :

$$\text{Average Precision @k}=\frac{\sum_{i=1}^{k} rel\,(i)P@i}{\sum_{i=1}^{n} rel\,(i)}$$

The mean average precision MAP@k at document cut off value k of a query set is the arithmetic mean of the average precisions of all queries in the query set.

4.2 Mean Reciprocal Rank

The Reciprocal Rank (RR) of the result set of query is defined as the reciprocal value of the highest ranking relevant document's rank in the result set[9]. For a query q, the rank positions of its relevant retrieved document is presented as r_1.Thus the Mean Reciprocal rank is represented as $1/r_1$

$$RR@k=\begin{cases}\frac{1}{i} if\ \exists i\ \leq k : rel(i)=1\ \wedge\ \forall j<i : rel(j)=0\\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise\end{cases}$$

We can define the Mean Reciprocal Rank of a query set is the average reciprocal rank of all queries in the query sets[8]. Let there are n number of queries in the query set. It is represented as

$$MRR=\frac{1}{n}\sum_{k=1}^{n} RR@k$$

4.3 Discounted Cumulative Gain

For the calculation of Discounted Cumulated Gain we first need to find the cumulated gain. In this relevance score of the given document is calculated as a gained value measure for its ranked position in the result set and the calculated gain is computed by summing from the position 1 to the last. Now the ranked document IDs are replaced by their relevance score thus ranked document lists become the gained value lists. Assuming the relevance scores 0 to 3 are used (3 denoting the highest value, 0 being the lowest). The average of the highest ranked value and the lowest ranked value is taken say it is σ_1.The average of σ_1 and the lowest ranked value is taken say it is σ_2. The average of σ_1 and the highest ranked value is taken say it is σ_3.The range of ranked values between σ_2 and the lowest ranked value has the relevance as 0. The range of ranked values between σ_2 and σ_1 has the relevance as 1. The range of ranked values between σ_1 and σ_3 has the relevance as 2. The range of ranked values between σ_3 and the highest ranked value has the relevance as 3.Let G[i] denotes the gain vector G at a position *i*. Cumulated Gain vector CG is defined recursively as[10]

$$CG[i]=\begin{cases}G[1], & if i=1\\ CG[i-1]+G[i], & otherwise\end{cases}$$

In the case of Discounted Cumulated Gain a discounting function is required that continuously and steadily reduces the document score when its rank increases. This is required to allow user persistence in evaluating the further documents. The most common way to perform this task is to divide the document score by the log of its rank. If b denotes the base of the logarithm, The Discounted Cumulated Gain is defined as follows [10]:

$$DCG[i] = \begin{cases} CG[i], & if i < b \\ DCG[i-1] + \dfrac{G[i]}{b_{\log i}}, & if i \geq b \end{cases}$$

4.4 Normalized Discounted Cumulative Gain

The normalised discounted Cumulated Gain nDCG@k of a query set is obtained by dividing the DCG@k of the result set rank ordered according to the scores by the DCG@k of the result set rank ordered according to the ideal scoring function. The ideal scoring function is the one that rank order results according to their ratings. This ideal scoring function is obtained by sorting documents of the result list according to the relevance thereby giving the maximum possible value of DCG till the position k and this DCG is known as Ideal Discounted Cumulative Gain IDCG@k[10]. For the given query the Normalised Discounted Cumulative Gain is given by:

$$nDCG@k = \frac{DCG@k}{IDCG@k}$$

5. Page Rank, HITS, SALSA and norm(P) algorithms

5.1 The neighborhood graph of the result set[12]

The algorithms discussed in this paper are based on topical endorsements. By topical endorsement we mean that the hyperlink on a given page u on a given topic endorses the authority of another page v with respect to given topic. Due to this topical endorsement it is economical to perform the link analysis on the hyperlinks that are endorsing each other. To perform this task effectively we need to construct the neighbourhood graph of the result set. For a given subset of documents a neighbourhood graph consists of all links between the documents. It consists of the documents that appear in the retrieved document set.

For this study let us consider a web graph G(V,E) having the vertex set V and edge set E ⊆ V X V .The result set of the query URLs is known as root set R ⊆ V. From this the algorithms construct a neighbourhood graph which consists of base set B ⊆ V (the root set and some of neighbouring vertices) and some of the edges in E are included in B. To define the neighbourhood graph formally sampling operator and link selection predicate are used.

For the given set A, the sampling operator $S_n[A]$ selects n elements uniformly at random from the set A.

$$\boldsymbol{S_n[A] = Aif \ \mid A \mid \leq n}$$

For the given set a link selection predicate P uses an edge (u,v) ∈ E. In this study three selection predicates are used.

$$all(u,v) \Leftrightarrow true$$
$$inter - host(u,v) \Leftrightarrow host(a) \neq host(b)$$

$$inter - domain(u, v) \Leftrightarrow domain(a) \neq domain(b)$$

Where host(u) represents the host of URL a and domain(u) represents the domain of URL u. Therefore *all* is true for all the links in the results set whereas *inter-host* is true for inter-host links and *inter-domain* is true for inter-domain links.

The outlink set O of the root set with respect to link selection predicate P is represented as:

$$\boldsymbol{O^P} = \bigcup_{\boldsymbol{u \in R}} \{ \boldsymbol{v} \in \boldsymbol{V} : (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} \wedge \boldsymbol{P}(\boldsymbol{u}, \boldsymbol{v}) \}$$

The outlink set O of the root set with respect to link selection predicate and sampling operator S is represented as:

$$\boldsymbol{O_s^P} = \bigcup_{\boldsymbol{u \in R}} \boldsymbol{S_S}[\{ \boldsymbol{v} \in \boldsymbol{V} : (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} \wedge \boldsymbol{P}(\boldsymbol{u}, \boldsymbol{v}) \}]$$

The inlink set I of the root set with respect to link selection predicate P is represented as:

$$\boldsymbol{I^P} = \bigcup_{\boldsymbol{v \in R}} \{ \boldsymbol{u} \in \boldsymbol{V} : (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} \wedge \boldsymbol{P}(\boldsymbol{u}, \boldsymbol{v}) \}$$

The inlink set I of the root set with respect to link selection predicate and sampling operator S is represented as:

$$\boldsymbol{I_s^P} = \bigcup_{\boldsymbol{v \in R}} \boldsymbol{S_s}[\{ \boldsymbol{u} \in \boldsymbol{V} : (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} \wedge \boldsymbol{P}\ (\boldsymbol{u}, \boldsymbol{v}) \}]$$

The base set $\boldsymbol{B_s^P}$ of the root set R with respect to P and s is defined as:

$$\boldsymbol{B_s^P} = \boldsymbol{R} \cup \boldsymbol{I_s^P} \cup \boldsymbol{O^P}$$

The neighbourhood graph ($\boldsymbol{B_s^P}$, $\boldsymbol{N_s^P}$) has the base set $\boldsymbol{B_s^P}$ as its vertex set and an edge set $\boldsymbol{N_s^P}$ containing those edges in E that are covered by $\boldsymbol{B_s^P}$ and permitted by P:

$$\boldsymbol{N_s^P} = \{ (\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{E} : \boldsymbol{u} \in \boldsymbol{B_s^P} \wedge \boldsymbol{v} \in \boldsymbol{B_s^P} \wedge \boldsymbol{P}(\boldsymbol{u}, \boldsymbol{v}) \}$$

5.2 PageRank Algorithm

PageRank is one of the most important ranking techniques used in today's search engines i.e. Google. PageRank is a simple, robust and reliable way to measure the importance of web pages, but it is also computationally advantageous in comparison to other ranking techniques in that it is query independent and content independent. The PageRank algorithm is based on the formula [3]:

$$pr(A) = (1 - d) + d \sum_{s \in I_A} \frac{pr(s)}{\deg(s)}$$

Where pr(s) is the PageRank of the vertex s

I_A is the in-neighbour set of the vertex Adeg(s) is the out-degree of vertex $s \sum_{s \in I_A} \frac{pr(s)}{\deg(s)}$ is the summation of PageRank of all the webpages pointed to given webpage.

Algorithm 1 PageRank algorithm(G,s,k)

```
d ← 0.85
n ← number of vertices in G
```

```
fori= 0 to n do
pr[i] ← s
end for
for j=0 to k do
for allpr[i] do
$pr_{in}$ ← sum of all incoming normalised PageRanks
pr[i] ← (1-d) + d($pr_{in}$)
  end for
end for
 avg ← sum(pr[i])/n
```

5.3 *HITS Algorithm*

In the HITS algorithm for a given query q a set of t highest ranked pages are selected. These pages are known as root set. From this base set S is constructed by including any page pointed to by a page and any page points to a page. For a given page i in S an authority score a(i) and hub score h(i) is assigned[4].

$$a(i) = \sum_{(j,i)\in E} h(j)$$

$$h(i) = \sum_{(i,j)\in E} a(j)$$

Algorithm 2 HITS algorithm(G,k)

```
G ← a collection of n linked pages
k ← a natural number
L ← adjacency matrix of the graph
Initialize $a_0$= $h_0$= (1,1,….1)
HITS –Iterate (G)
Let z denote the vector(1,1,1,1,….1) $\in R^n$
$a_0 \leftarrow h_0 \leftarrow z$
$k \leftarrow 1$
Repeat
Apply the I operation to $(a_{k-1}, h_{k-1})$ as follows
    $a_k \leftarrow L^T L a_{k-1}$
Apply the O operation to $(a_{k-1}, h_{k-1})$ as follows
                              $h_k \leftarrow L L^T h_{k-1}$
$a_k \leftarrow \frac{a_k}{\|a_k\|}$          //normalization
$h_k \leftarrow \frac{h_k}{\|h_k\|}$          //normalization
                              $k \leftarrow k + 1$
until$|a_k - a_{k-1}| < E_a$ and $|h_k - h_{k-1}| < E_h$
```

```
    **return** $a_k$ and $h_k$
```

5.4 *SALSA Algorithm*

The Stochastic Approach for Link Structure Analysis is based on the concept of Markov Chains and uses the stochastic properties of random walks which is performed on the collection of webpages. In this approach a neighbourhood graph is determined first. On that neighbourhood graph one step backward and one step forward random walk is performed. The stationary probability distribution gives the authority score of SALSA algorithm.

The Hub matrix H [5] defined as follows:

$$h_{i,j} = \sum_{\{k|(i_h,k_a),(j_h,k_a \in G\}} \frac{1}{\deg(i_h)} \frac{1}{\deg(k_a)}$$

The Authority matrix A [5] defined as follows:

$$a_{i,j} = \sum_{\{k|(k_h,i_a),(k_h,j_a \in G\}} \frac{1}{\deg(i_a)} \frac{1}{\deg(k_h)}$$

Algorithm 3 SALSA Algorithm(G,s,k)

```
B ← neighbourhood graph
    $B = \bigcup_{u \in R} \{u\} \cup S_m[\{v \in V : (u,v) \in E\}] \cup S_n[\{\in V : (u,v) \in E\}]$
B^A = { u ∈ B : in(u) >0}
**For** all u ∈ B **do**
    $A(u) = \begin{cases} \frac{1}{|B^A|} & if\ u \in B^A \\ 0 & otherwise \end{cases}$
**Repeat** until A converges:
**For all** $u \in B^A$ **do**
    $A'(u) = \sum_{(v,u) \in N} \sum_{(v,w) \in N} \frac{A(w)}{out(v) in(w)}$
**For all** $u \in B^A$ **do** A(u) = $A'(u)$
```

5.5 *Norm(P) Algorithm*

It works on the principle that small authority weights should contribute less to computation of the hub weights[7].

Algorithm 4 Norm(P) algorithm(G,s,k)

```
**Repeat until convergence**
**O operation:** hubs compute the p-norm of the authority weight vector
```

$$h_i = (\sum_{j:i\to j} a_j^p)^{\frac{1}{p}} = \|F(i)\|_p$$

3: **I operation:** authorities collect the hub weights

$$a_i = \sum_{j:j\to i} h_j$$

4: **Normalise** weights under some norm

6. Experimental Results

For illustrating the results obtained by comparing the MAP,MRR and NDCG values between the present scenario of considering the link between the two given pages and considering all the outlinks on a given page while making the adjacency matrix. In the bar graphs given below we use crisp values to define the present scenario and conditional probability to define the scenario where all the outlinks on a given page are considered for making the adjacency matric for the given web graph which is given as an input to the link based algorithms.From the experiments and the bar graphs presented above we can conclude that in some algorithms with conditional probability taken into account these perform way too better than their corresponding counterpart of crisp values.

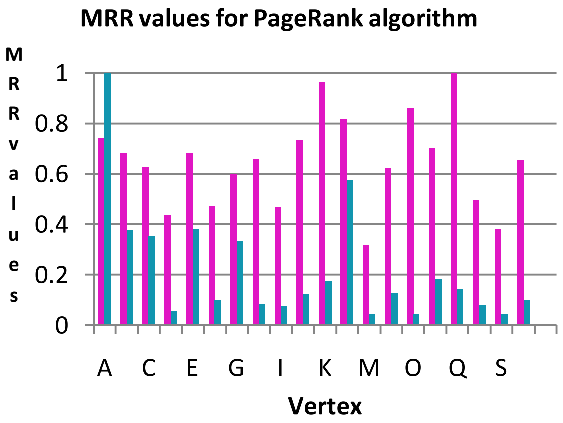

Figure 1

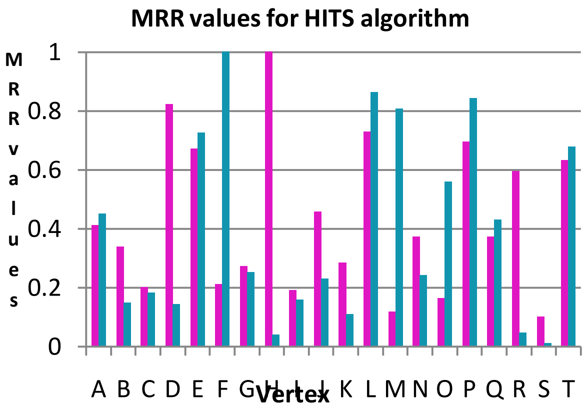

Figure 2

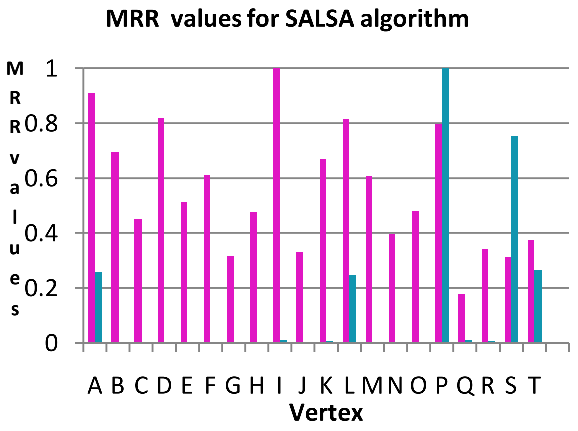

Figure 3

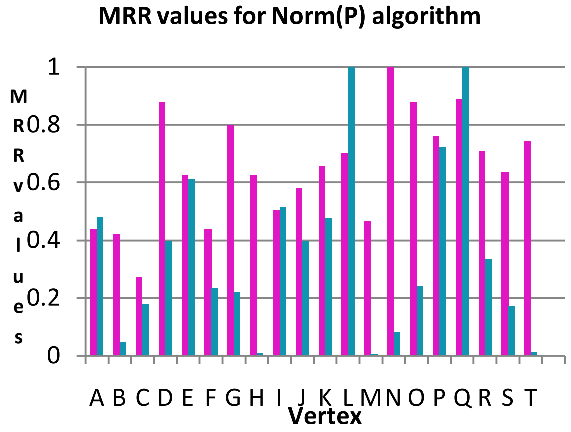

Figure 4

In the figure 1 though in most of the cases crisp values perform better in the MRR values as compared to the conditional probability. In figure 2 the MRR values of HITS algorithms the performance of conditional probability values is almost similar to crisp values. Here the values are conditional probability when taken as whole is slightly less than the crisp values taken as whole. In figure 3 MRR values of the crisp values clearly dominate the MRR values of the conditional probability. In figure 4 the MRR values of Norm (P) algorithm the results with the crisp values are better as compared to conditional probability.

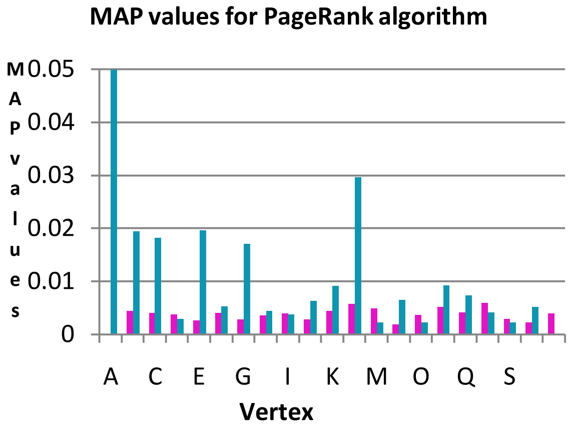

Figure 5

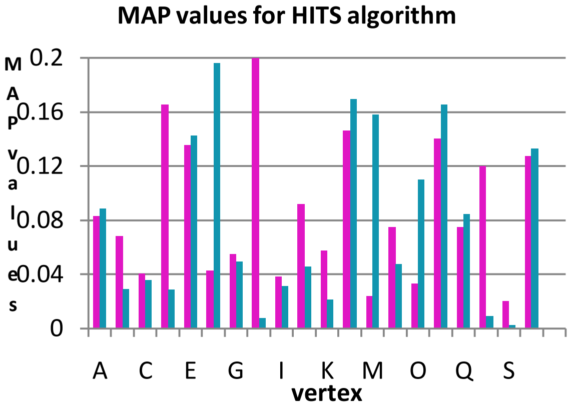

Figure 6

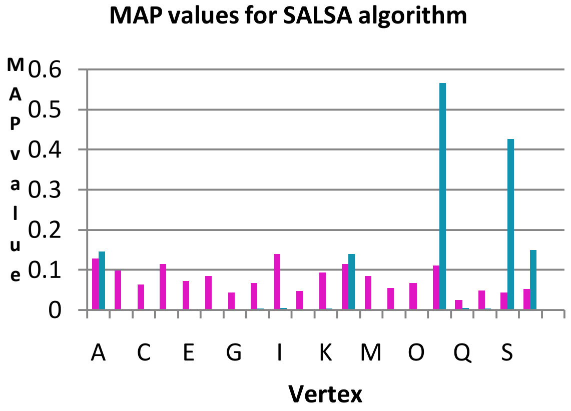

Figure 7

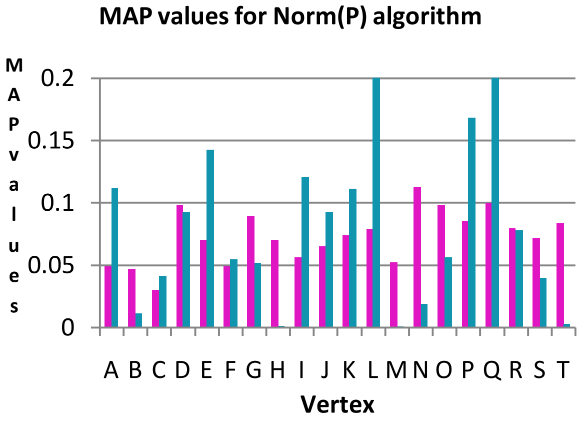

Figure 8

In figure 5 the performance of the MAP values of the PageRank algorithm are far better than the values obtained with crisp values. In figure 6 the MAP values for crisp values and conditional probability of HITS algorithms are compared. The values obtained in the case of crisp values are almost similar to the values obtained in the case of conditional probability. In figure 7 the MAP values for SALSA algorithm are compared. Though the higher values corresponding to conditional probability are smaller in number but when these are taken as a whole these are marginally higher than the values obtained in the case of crisp values. In figure 8 the results obtained in the case of MAP value for conditional probability are more efficient than the corresponding crisp values in the case of Norm (P) algorithm.

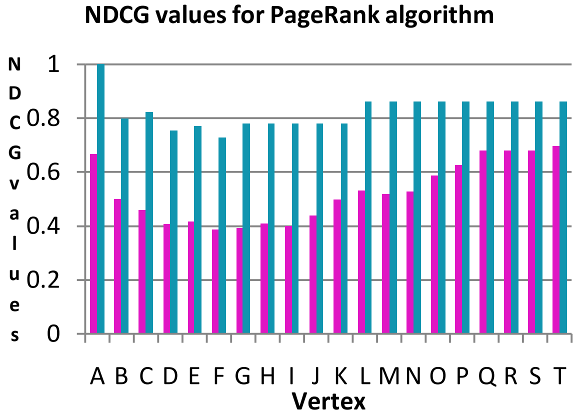

Figure 9

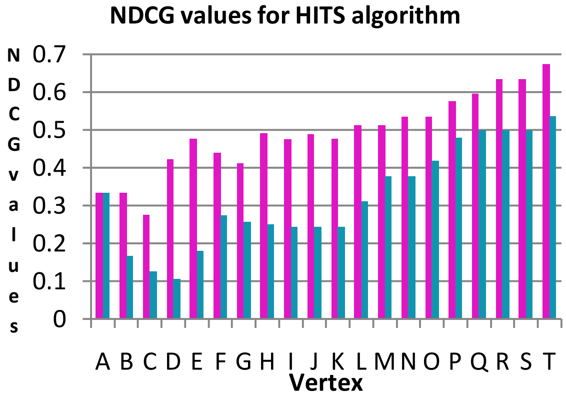

Figure 10

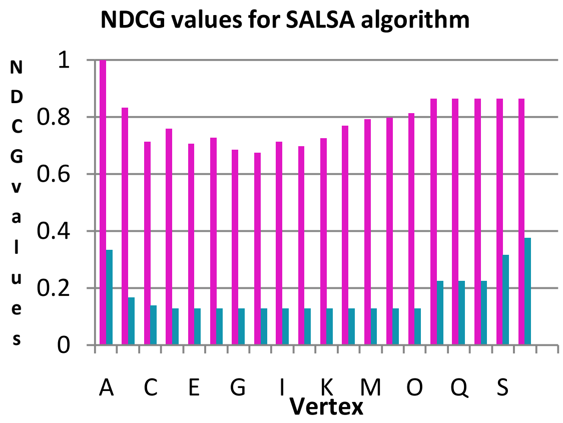

Figure 11

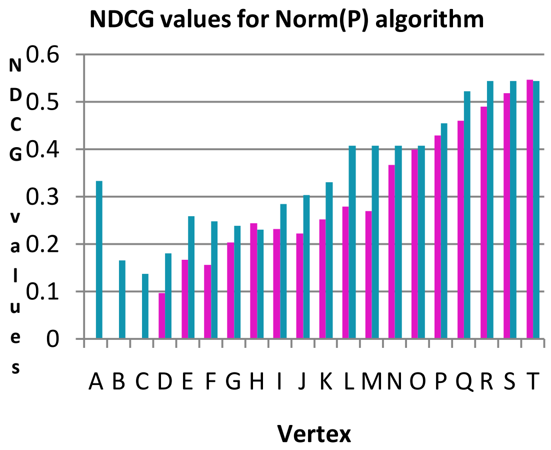

Figure 12

In the figure 9 the NDCG value of the PageRank algorithm are compared. In this case the NDCG values for the conditional probability are relatively higher than their corresponding crisp values. In the figure 10 the performance for the NDCG values in the case of crisp values are better than the values of the conditional probability in the HITS algorithm. In the figure 11the NDCG values for SALSA algorithm of the crisp values are comparatively higher than their corresponding values in the case of conditional probability. In figure 12 NDCG values of the conditional probability clearly dominate the NDCG crisp values of the in the Norm (P) algorithm.

6. Conclusion

With this study we propose that instead of considering only the concerned outlink on a given page we should consider all the outlinks .Based on these outlinks we should calculate the probability of visiting the given outlink and this probability is used in the adjacency matrix of the web graph. In the case of MRR values the performance these probability values calculations are not very sound as compared to crisp values. The performance of MAP values is considered to be same in both the scenario. For the NDCG values PageRank algorithm and Norm (P) algorithm perform exceptionally well in case of conditional probability.

References

[1] MONIKA HENZINGER, Link Analysis in Web Information Retrieval, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000.

[2] B. J. JANSEN, T. SARACEVIC, J. BATEMAN AND A. SPINK Real life information retrieval: a study of user queries on the web. ACM SIGIR Forum, 1998.

[3] S.BRIN ,L.PAGE, The anatomy of a large-scale hypertextual web search engine, Proceedings of the 7th International World Wide web Conference, 1998.

[4] J.KLEINBERG, Authoritative sources in a hyperlinked environment, Journal of ACM (JASM), 1999.

[5] R.LEMPEL, S.MORAN, The stochastic approach for link-structure analysis (SALSA) and the TKC effect, Proceedings of the 9th International World Wide web Conference, 2000.

[6] R. LEMPEL AND S. MORAN SALSA : The stochastic approach for link structure analysis,ACM Transaction on Informattion Systems, 2001.

[7] CYNTHIA RUDIN, Ranking with a P-Norm Push ,Springer-Verlag Berlin Heidelberg, 2006.

[8] TIE-YAN LIU, Learning to Rank for Information Retrieval,ACM SIGIR, 2008.

[9] E.M. VOORHEES , TREC-8 Question Answering Track Report,Proceedings of the 8th Text Retrieval Conference National Institute of Standards and Technology (NIST), 1999.

[10] KALERVO JARVELIN AND JAANA KEKALAINE ,Cumulated Gain-Based Evaluation of IR Techniques, ACM Transactions on Information Systems, Vol. 20, No. 4, Pages 422–446, October 2002.

[11] A.BORODIN, G.O.ROBERTS, J.S.ROSENTHAL, P.TSAPARAS, Link analysis ranking: algorithms, theory, and experiments, ACM Transactions on Internet Technology, 2005.

[12] MARC NAJORK,Comparing the Effectiveness of HITS and SALSA,Proceedings of the sixteenth ACM Conference on information and knowledge management, 2007.