Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An efficient page ranking approach based on vector norms using *s*Norm(*p*) algorithm

Shubham Goel[a], Ravinder Kumar[a], Munish Kumar[b,*], Vikram Chopra[c]

[a] Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, Patiala, India
[b] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, India
[c] Electrical & Instrumentation Engineering Department, Thapar Institute of Engineering & Technology, Patiala, India

ARTICLE INFO

Keywords:
Page ranking
HITS
SALSA
Vector Norm
Mean reciprocal rank
p-Norm

ABSTRACT

In the whole world, the internet is exercised by millions of people every day for information retrieval. Even for a small to smaller task like fixing a fan, to cook food or even to iron clothes persons opt to search the web. To fulfill the information needs of people, there are billions of web pages, each having a different degree of relevance to the topic of interest (TOI), scattered throughout the web but this huge size makes manual information retrieval impossible. The page ranking algorithm is an integral part of search engines as it arranges web pages associated with a queried TOI in order of their relevance level. It, therefore, plays an important role in regulating the search quality and user experience for information retrieval. PageRank, HITS, and SALSA are well-known page ranking algorithm based on link structure analysis of a seed set, but ranking given by them has not yet been efficient. In this paper, we propose a variant of SALSA to give sNorm(p) for the efficient ranking of web pages. Our approach relies on a p-Norm from Vector Norm family in a novel way for the ranking of web pages as Vector Norms can reduce the impact of low authority weight in hub weight calculation in an efficient way. Our study, then compares the rankings given by PageRank, HITS, SALSA, and sNorm(p) to the same pages in the same query. The effectiveness of the proposed approach over state of the art methods has been shown using performance measurement technique, Mean Reciprocal Rank (MRR), Precision, Mean Average Precision (MAP), Discounted Cumulative Gain (DCG) and Normalized DCG (NDCG). The experimentation is performed on a dataset acquired after pre-processing of the results collected from initial few pages retrieved for a query by the Google search engine. Based on the type and amount of in-hand domain expertise 30 queries are designed. The extensive evaluation and result analysis are performed using MRR, Precision@k, MAP, DCG, and NDCG as the performance measuring statistical metrics. Furthermore, results are statistically verified using a significance test. Findings show that our approach outperforms state of the art methods by attaining 0.8666 as MRR value, 0.7957 as MAP value. Thus contributing to the improvement in the ranking of web pages more efficiently as compared to its counterparts.

1. Introduction

The enormous volume and unstructured nature of information over the web are causing a big threat to the information retrieval efficiency of search engines. Thousands of results are returned by search engines for a search query out of which only some are

[*] Corresponding author.
E-mail addresses: ravinder@thapar.edu (R. Kumar), munishcse@gmail.com (M. Kumar), vikram.chopra@thapar.edu (V. Chopra).

https://doi.org/10.1016/j.ipm.2019.02.004
Received 6 July 2018; Received in revised form 7 February 2019; Accepted 9 February 2019
Available online 01 March 2019

actually relevant to the topic of interest (TOI). For most of the users, retrieval of some most relevant results is more than enough to acquiescent them and rest all retrieved results is a waste for them. Two classical problems of information retrieval, polysemy, and synonymy, also contributed in degrading the efficiency of search engines by an increase in a number of search results of varying degree of relevance. Due to this, search sessions are getting longer in searching for relevant information and these results in the increase of user's dissatisfaction level. They start perceiving relevant information retrieval as the difficult task (Liu, Kim, & Creel, 2015).

Generally, for information retrieval, every search engine uses content similarity matching concept to match, similarity between a query and a web page. In primitive search engines, the ranking is performed on the basis of strength of content similarity match calculated by counting the number of time a keyword presents in a web page. With the advancement of technology, content-based spamming becomes a popular means of fraudulent ranking of spammer web pages which result in failure of content-based ranking and the emergence of a new era of link-based ranking.

Together with the textual content of web pages, link structure present between the web pages is very informative. So, with proper analysis of link structure information on web pages, most relevant, authoritative resources can be arranged in first few result pages in-order of their relevancy level. PageRank (Brin & Page, 1998), HITS (Kleinberg, 1999), and SALSA (Lempel & Moran, 2000) are a few well-known link structures based page ranking algorithms. To reduce task difficulty in Information Retrieval (IR) by search engines, more and more researchers are studying in the field of IR (Langville & Meyer, 2011, 2012). Some of them work on query optimization, click-through logs and some of redesigning page ranking algorithms as presented in the next section. But before designing any solution, differences and overlap in a web search engine must be studied (Spink, Jansen, Blakely, & Koshman, 2006) i.e. information retrieval and ranking algorithms used by different web search engines. As per our study efficient ranking of search results in order of their relevancy level to the topic of interest (TOI) can be an effective solution to problems of IR with search engines. In this paper, we propose to employ the concept of Vector Norms (Householder, 2013; Schott, 2016; Wilkinson, 1965) to improve search engine performance and reducing the number of results returned to a user by efficiently ranking the retrieved web pages in order of their relevancy level to topic of interest (TOI). Correct sequencing will result in improvement of user satisfaction level and search quality.

In Section 2, of this paper, a short description of the work carried out by other researchers is presented. In Section 3, objectives of our research work are presented. Section 4, discuss basics of link analysis and already existing system model. Proposed methodology has been presented in Section 5. Section 6 discusses data set acquisition and pre-processing performed on collecting corpus. In Section 7, we have presented experimental results based on the proposed technique and discussion of the effectiveness of the proposed technique. Comparisons with baselines are described in Section 8. Finally, inferences and future directions are outlined in Section 9.

2. Related work

Different surveys pointed out that search results obtained by different web search engines do not relate to Topic of Interest (TOI) and exhibit a low degree of user satisfaction. Jansen and Molina (2006) evaluated the relevancy of search results for e-commerce links retrieved by search engines, namely, Excite, Froogle, Google, Overture and Yahoo directories. Each search engine belongs to a different category and findings reveal that the search result obtained by e-commerce search engines like Froogle is more relevant than the result obtained by other engines.

Satisfaction of information need of the user is the main motive behind the development of web search engines as manual search gets impossible due to increase in the size of the web. Spink et al. (2006) measured the overlap in the results obtained by four web search engines, namely, Google, Yahoo, MSN and ASK. Results on an only first page for a large set of queries are considered for the experiment. Jansen and Rieh (2010) identified and compared 17 important theoretical constructs in the field of information searching & retrieval. Their results reveal that trade-off of these constructs is very important for efficient working of an information system.

2.1. Ranking algorithms

Ranking algorithms aim at the ranking of web pages in order of their relevancy level w.r.t TOI. Brin and Page (1998) have given a famous PageRank algorithm which later forms an integral part of Google search engine. In PageRank algorithm more the number of inward links to a web page from other web pages in a web graph, higher is the rank of a web page. Term-based searching is employed to retrieval of information from the web and then ranking is performed using PageRank algorithm. With the growth of web and e-commerce applications, every business organization wants their web page to be ranked higher in search engines. Many professionals start researching PageRank algorithm and using its loopholes to take benefits (Langville & Meyer, 2011; Langville & Meyer, 2012). Many defects in the classical PageRank algorithm, namely link, and keyword spamming are discussed by Kumar, Singh, and Mohan (2016) and Yang and Zheng (2016). Thalhammer and Rettinger (2016) have had done an analysis of the PageRank algorithm on a link graph of documents on Wikipedia. Impact of link position and context in document ranking is studied and the results are compared with the ranking given by the classical PageRank algorithm. These tricks can rank a web page on the higher rank, but actually this web page has nothing informative about TOI. Improvements of the classical PageRank algorithm are done many researchers. Shen, Huang, Carpentieri, Gu, and Wen (2017) analyzed similar link distribution problem of PageRank algorithm and proposed a solution to it by eliminating identical sub-rows in the transition matrix. An elimination operator is generated to transform the problem into an equivalent, but a sparse problem. The results obtained reveal that there is a decrease in density and computation cost of solving PageRank problems. Guha, Kundu, and Duttagupta (2015) have given a link based weighing system for ranking of web pages. The number of links available on a web page will decide the weighing factor. The rank of a web page is subsequently multiplied by the weighing factor. Yang and Zheng (2016) proposed an improved PageRank algorithm by adding link similarity and time feedback features to the classical one. Kleinberg (1999) devised Hyperlink-Induced Topic Search (HITS) and currently, it is an integral part of the Ask search engine. In HITS, the set of web pages is structured in authority and hub resources. The resource is called authority if it contains significant information about the queried topic and hub if it only advertises authority resources or indirectly provides information about the query topic. Therefore, two types of scores are assigned to each web page: a hub score and an authority score. Hub score is the sum of the authority score of those web pages which are advertised by the hub. Authority score is the sum of the hub score of those web pages which are advertised that authoritative web page. Both hub and authority scores are updated alternatively until equilibrium is reached. The ranking of web pages is performed on the basis of authority score trade-off. Najork, Zaragoza, and Taylor (2007) have done the evaluation of HITS algorithm and compared its effectiveness with PageRank and other state of the art methods for making use of anchor text. Many Improvements of HITS algorithm are also suggested by researchers. Borodin, Roberts, Rosenthal, and Tsaparas (2005) devised Hub-Threshold HITS algorithm by not considering the effect of medial authorities based on the fact that they are rarely clicked by a user in the real search scenario. Only the web pages corresponding to the highest K authorities are considered for hub score calculation, but the procedure for authority score calculation remains unchanged. Here K is an algorithm parameter. Yang (2016) has given an improved version of the HITS algorithm by considering both link structure analysis and content similarity for the ranking of web pages.

Lempel and Moran (2000), proposed Stochastic Approach for Link-Structure Analysis (SALSA) which replaces mutual reinforcement approach of Kleinberg. Coupling between authority and hubs is less tight in SALSA as compared to HITS. The whole corpus of web pages related to a query topic is represented by a bipartite graph with the hub and authority as two sides of a graph. Theory of Markov chains is used in SALSA with stochastic properties of random walks on the bipartite graph of web pages for creation on transition matrices. In the conventional scenario, the single Markov chain is used by a search engine like Google, but in SALSA two Markov chains are used, one representing the backward link and another representing the forward link. Two transition matrices are used to represent each Markov chain respectively. Ranking of authoritative web pages needs to be done according to probabilities identified with the trade-off of two transition matrices. Tagarelli, Kabbur, and Karypis (2015) discussed information retrieval methods, page ranking algorithms and features of web pages on which the ranking of pages can be performed. Singh, Sharma, Rishi, and Akhtar (2017) discussed the problem faced by web miners and different algorithms used for page ranking w.r.t applications in the field of e-commerce for the betterment of business. Li, Zhao, and Garcia-Molina (2012) have given a path based, Hierarchical Navigation Path (HNP) and PathRank algorithms for web page ranking. For website browsing HNP provides a multi-step navigational information and hints about the content of the destination page for the visitor. Most of the links only provide a recommendation, but in the case of websites, it is only for navigation purpose. Singh and Sharma (2013) have given Enhanced-Ratio Rank which considers a web page to be more important if incoming links to that page are visited by more users in comparison to other pages. Weight of both incoming and outgoing links is taken in a defined proportion for a popular page. Zin, Tin, Hama, and Toriu (2015) used a concept of queuing theory of operation research and stochastic water storage theory in hydrology to design an approach for web page ranking. Gupta, Dixit, and Devi (2015) developed a crawling technique only to download the relevant documents. The ranking of the downloaded documents is then performed on the basis of synonyms, user domain profiles and other information, i.e. average amount of time spent by past users. Campos, Dias, Jorge, and Nunes (2016) proposed GTE-Rank, to regulate the ranking of web pages retrieved by a search engine to answer the time sensitive queries. The temporal features of a web page are extracted using a content-based approach to information retrieval. It is a combination of topical and temporal rank values. Results are better than baseline methods when tested for various evolutionary parameters. A helpful summary of the referred article is provided by the text around a citation mark. Doslu and Bingol, 2016) analyzed the impact of ranking, the articles related to a TOI on the basis of citation context by applying different ranking algorithms on the directed network of articles created by citation context. Negm, AbdelRahman, and Bahgat (2017) proposed PREFCA, a Portal Retrieval Engine based on Formal Concept Analysis that considers portal as the main search unit instead of a web page. Semantic relations between portal and web pages are maintained for efficient information retrieval. F-measure accuracy and MAP values for PREFCA are preferable to other search engines that use approaches, namely Latent Semantic Analysis (LSA), TF-IDF and BM25. In summary, some improvement in web search engine performance in terms of ranking relevant documents was obtained by previous studies but that too at cost of complex computations.

SALSA is a strong baseline in terms of effectiveness because it combines key thoughts from both HITS and PageRank. SALSA uses a query specific neighborhood graph similar to the one as HITS does, and it also computes a hub score and an authority score for each node in the neighborhood graph. However, while HITS uses a methodology called “mutual enforcement” where hubs enforce authorities and vice versa, SALSA processes these scores by performing two independent random walks on the neighborhood graph, a hub walk and an authority walk, thus embracing a key thought of PageRank. In this research, we propose a variant of the well-known SALSA algorithm for efficient ranking of web pages without much increase in computations. Vector Norms family is used to devise a modified system, namely, sNorm(p) to perform the ranking of web pages on the basis of the link structure analysis of web pages in a hyperlinked environment.

3. Research objectives

The goal of our research is to devise a web page ranking algorithm to efficiently re-rank the web search results retrieved by a web search engine in-order of their relevancy level. The specific research objectives of the study are as follows:

- To pre-process the web graph constructed based on link structure prevailing among the web pages present on initial few result

Table 1
Summary of Symbols and Notation.

Symbols	Description	Symbols	Description
TOI	Topic of interest	$\overrightarrow{H}$	Hub vector
C	Collection of web pages & their link structure	a_i	Authority score of *i*th node
G	Graph for C, nodes as web page & edges as link structure	h_i	Hub score of *i*th node
M	Adjacency matrix for G	$B(i)$	Set of nodes pointing to *i*th node
M^T	Transpose of M	$F(i)$	Set of nodes pointed by *i*th node
$\overrightarrow{V}$	Page rank vector	N_h	Set of nodes on hub side of bipartite graph
n	Number of nodes in G	N_a	Set of nodes on authority side of bipartite graph
$\overrightarrow{A}$	Authority vector	p	p-norm operation

pages retrieved by a web search engine to generate a seed set for the query.
- To design and implement sNorm(p) in order to rank web pages in the seed set.
- To compare ranking hierarchy assigned by PageRank, HITS, SALSA, and sNorm(p) to web pages in the seed set.
- To compute MRR, Precision@k, MAP, DCG, and NDCG as evaluation parameters for quantization the efficiency of ranking algorithms.
- To verify the obtained results using a significance test.

4. Problem formulation

Problem researched in this section had aroused with the growth of the information network on a web. Numerous web pages are created to provide detailed information to users but at the cost of increase in search complexity for retrieval of relevant and credible web pages with respect to TOI by a search engine. In order to tackle this problem, efficient systems need to be designed which can rank the web pages relevant to TOI according to their relevancy level. To predict the ranking of web pages, merely a textual content similarity is not enough as used in the past decade as a solution to the above-mentioned problem. Therefore, aroused a demand for something new and more efficient. The power of text-based search and a ranking algorithm can be improved, if the link structure information on web pages is embedded in it. The link structure information present between the web pages in a form of hyperlinks can prove to be informative in predicting authenticated and relevant web pages w.r.t. query. So the contribution of both textual and link structure information on web pages is required for the efficient ranking of web pages. Web communities have also made it a standard practice that must be followed by any web page ranking algorithm. Many successful systems were developed based on the above premise. Web standards are continuously going to evolve with market demand to give better performance. A brief summary of symbols and notations used in this work is depicted in Table 1.

Definition 1: [Web Page (WP)]: Let $\mathrm{L_w}$, $\mathrm{L_h}$ be respectively the sets of web pages retrieved, hyperlinks present in the corresponding web pages. For example, $p \rightarrow q$ considering p and q as web pages and " → " as a hyperlink present in p to link it with q. Together $(p, \rightarrow) \in WP$ signifies that author of page p recommending page q to the user visiting p.

Definition 2: [Equilibrium and Iteration]: Equilibrium is the state in which there is no further change in values of a vector or matrices, in spite of any number of further iterations after equilibrium is reached. For example, $\overrightarrow{\mathrm{V}}_1^1 \cong \overrightarrow{\mathrm{V}}_1^k$, where $\overrightarrow{\mathrm{V}}_1^1$ depicts the value of the vector at an equilibrium state and $\overrightarrow{\mathrm{V}}_1^k$ depicts the value of the vector after k iterations.

4.1. PageRank

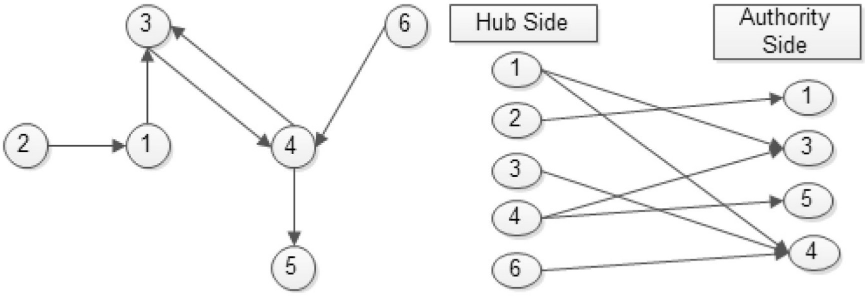
PageRank was the first designed system to rank web pages and became a basis for every other ranking algorithms designed afterward. The entire network of web pages relevant to TOI is represented by a directed graph. At every node, importance is equally distributed among all outgoing links or edges. An adjacency matrix M is created in the graph considering only weight on incoming edges at each node from other nodes. Initially, each element of page rank vector $\overrightarrow{V}$ is assumed to be same and equal to $1/n$, where n is a number of nodes in a graph. Page rank vector is updated recursively using Eq. (1) until equilibrium is attained.

$$\overrightarrow{V_1} = M\overrightarrow{V},\ \overrightarrow{V_2} = M^2\overrightarrow{V},\ \ldots\ldots\ldots\ldots\ldots\ldots,\overrightarrow{V_k} = M^k\overrightarrow{V} \tag{1}$$

$\overrightarrow{V_k}$ is the final page rank vector and ranking of web pages are done on the basis of it. This PageRank algorithm is used by the Google search engine but still ranking obtained is not efficient.

4.2. Hyper-Induced Topic Search (HITS)

In HITS algorithm, web pages obtained by a search engine based on PageRank algorithm are used as a corpus for re-ranking. The idea is to classify corpus of web pages in qualitative authorities and hubs is put forward in HITS. Hub and authority score for each web page in a corpus is computed. Iterative vector for qualitative hub page ($\overrightarrow{H}$) and authority page ($\overrightarrow{A}$) are computed for nodes of the web graph using equations (Eqs. (2) and (3)).



(a) (b)

Fig. 1. Illustration of (a) web graph, (b) corresponding bipartite graph.

$$\vec{A} = M^T.\vec{H} \tag{2}$$

$$\vec{H} = M.\vec{A} \tag{3}$$

Either initialize each element of $\vec{A} = 1$ or $\vec{H} = 1$. The value of vectors after k number of steps $\vec{A_k}$ and $\vec{H_k}$ i.e. after equilibrium state is reached will represent the final hub and authority scores of each web page. The ranking is performed on the basis of vector $\vec{A_k}$. The vectors $\vec{A_k}$ and $\vec{H_k}$ converge to a principal right eigenvector of matrices W_1 and W_2 respectively, shown in equations (Eqs. (4) and (5)).

$$W_1 = M^TM \tag{4}$$

$$W_2 = MM^T \tag{5}$$

4.3. Stochastic Approach for Link-Structure Analysis (SALSA)

In SALSA algorithm, the scheme of PageRank and HITS algorithms is mixed. Theory of the random walk is combined with the idea of a hub and authoritative web pages. Bipartite graph (G_b) is created from the Graph (G), obtained from the list of web pages retrieved by the PageRank algorithm, for illustration as a given graph G and it is bipartite graph G_b as shown in Fig. 1(a) and (b), respectively. One side of the bipartite graph is called hub side and the other side is called authority side. Each web page in a collection can be present on both sides.

Random walks have been performed to identify highly reachable nodes in a graph, under the principle that is the highly reachable node is visited more frequently as compared to others. In SALSA, two Random walks have been conducted on a bipartite graph, one starting from authority side and another from the hub side by picking up node randomly. Two transition matrices, Hub (H) and Authority (A) are computed corresponding to two Markov chains one for each walk. The adjacency matrix (M), for a bipartite graph where each node is a web page and edge set, represents its hyperlink structure. Matrix (M_r) computed by dividing each non-zero entry of M by the sum of the entries of its row and matrix M_c, computed by dividing each non-zero entry of M by the sum of the entries in its column. Matrices H and A are computed using equations (Eqs. (6) and (7)).

$$A = M_c{}^TM_r \tag{6}$$

$$H = M_rM_c{}^T \tag{7}$$

The principal Eigenvector of H and A, considering only non-zero rows and columns will give a final authority vector $\vec{A}$ and hub vector $\vec{H}$. The ranking of web pages is done on the basis of the values of $\vec{A}$. This algorithm can be thought of as a variation of the HITS algorithm. Update operation of each authority a_i and the hub h_i score of a web page is shown in equations (Eqs. (8) and (9)).

$$a_i = \sum_{j:j\in B(i)} \frac{1}{|F(j)|}h_j \tag{8}$$

$$h_i = \sum_{j:j\in F(i)} \frac{1}{|B(j)|}a_j \tag{9}$$

5. Proposed model

We have proposed an efficient algorithm to merge schemes of p-norm and SALSA, to give sNorm(p). Theory of p-norm from the family of vector norms is embedded with SALSA. Based on the premise that higher authority weight will result in higher hub weight or vice-versa, the effect of higher authority weight must be increased, whereas low authority weight must be decreased while calculating hub weight. Proper scaling of the weight must be performed to reduce the contribution of low authority weight in hub

Algorithm 1
*s*Norm(*p*): Optimized web page ranking algorithm.

```
1 **Input:** node set $N_{h,}$ node set $N_{a,}$p-norm value $p$
2 **Output:** Authority Vector $\overrightarrow{A}$, Hub Vector $\overrightarrow{H}$
3 Initialize $\overrightarrow{A} = 1$ [where each element of $\overrightarrow{A}$ is member of $N_a$]
4 **Repeat** [Step 5 – 12] Until weights converges
5 *for* $i \in N_h$ do
6   | *for* j $\in F(i)$ do
7   |   | $temp = temp + \frac{a_j{}^p}{|B(j)|}$
8   | $h_i = \sqrt[p]{temp}$
9 *for* $k \in N_a$ do
10  | *for* $l \in B(k)$ do
11  |   | $temp = temp + \frac{h_l{}^p}{|F(l)|}$
12  | $a_k = \sqrt[p]{temp}$
13 Update $\overrightarrow{A}$ with new values of authority weights of its elements.
14 Normalize $\overrightarrow{A}$
```

weight calculation or vice-versa. To implement above idea p-norm is used. Initially, a graph G is created from the elements and hyperlink structure of elements of seed set R_f where each element corresponds to a web page. The bipartite graph (G_b) is created for a graph (G) similarly graph as in Fig. 1(b) is created from the graph in Fig. 1(a). Both sides of a bipartite graph can have duplicate nodes on different sides, but not on the same side. As shown in Fig. 1(b), nodes 1,2,3,4 and 6 are existing in hub side and correspondingly nodes 1, 3 and 4 are existing in authority side also. But, these nodes can't be duplicated in hub side or authority side. Duplicity feature is introduced to make a same web page behave authoritative resource at one time and hub resource at another time. Similar to SALSA two random walks are performed on a bipartite graph to find highly reachable nodes in a bipartite graph. In each step, the two edges of a bipartite graph are traversed by visiting nodes from either authority side or hub side of a graph in each walk. Both walks will start from the different sides of a bipartite graph and each path of length two in the bipartite graph will represent a traversal of one edge from authority to hub and moves backward along with an edge from hub to authority. Both walks will correspond to a separate Markov chain, i.e. hub chain and authority chain, depending on direction on a visit during a random walk. Required steps for proposed model are presented in Algorithm 1.

We set up the authority weight of the *i*th node to be p-norm of the hub vector of nodes pointing to the *i*th node. Hub weight of the *i*th node are the p-norm of the vector of authority weight of those nodes which are pointed to by the *i*th node. Here, the node means a corresponding web page. Each authority a_i and hub h_i score of web pages are updated according to the algorithm.

The ranking of web pages is performed on the basis of authority weights of the nodes after completion of the algorithm. Here nodes refer to web pages. Our idea will bring a new aspect of the solution to the problem of link analysis in the hyperlinked environment, which is defined as the process of finding web pages highly relevant to a query initiated by a user. In the stochastic models discussed in Section 4, the transition coefficients associated with the link from X to Y, in PageRank a function of out-degree (X); in HITS a constant value; and in SALSA either a function of out-degree(X) or a function of in-degree(Y), depending on the direction of the transition. In all three, all the out links from X or all the in links to Y are evaluated equally. The modifications that we propose, by contrast, give different normalized weights to different links, so as to avoid or alleviate the effect of web local aggregation. Our proposed algorithm evaluation can only be derived from the web graph structure, not query or topic basis.

6. Data acquisition and pre-processing

In actual scenario, the whole web graph is like a giant mesh of web pages, interrelated by millions and millions of hyperlinks. We have concentrated only on some parts of this giant mesh as it is not possible to work with the entire web graph for measuring the performance of the proposed system model. Based on the type and amount of in-hand domain expertise, 30 queries or TOI are selected. For a given TOI in a search query, collection C of web pages is created. It will get a mixture of both authority and hub web pages in some random proportion. Following steps are used to create collection C.

- Initially using a term based search engine like Google, a query Q is formulated for a TOI, a set R called root set is collected. Search results from initial few pages of results obtained by a search engine are the elements of set R.
- Set R of web pages is very heterogeneous and generally, there will be a few links from one web page to another. So, graph obtained by linking web pages based on hyperlinks is nearly disconnected. Algorithms to find relevant authorities and hubs cannot be enforced on a disconnected graph as it will result in a highly sparse matrix. Moreover, there will be very less chance for most relevant authority resources to be listed in a set R due to keyword spamming and synonymy problems. However, they are most supposed to be recommended by at least one of the web pages in the set R or by using hyperlinks to those authoritative resources.
- So the extension of set R is done by including all web pages which are recommended by every web page in set R and web pages which are recommending at least one web page in set R. Extended set R is denoted by R_{ex}.
- Set R_{ex} is shown by a multi-node directed graph where each node corresponds to an element of set R_{ex} and each edge corresponds

to a hyperlink present between elements of R_{ex}. Some of the hyperlinks present between elements of set R_{ex} are not informative such as intra-domain links, links created on the basis of mutual agreement and links to pop up advertisements. Intra-domain hyperlinks are used to join different sections of the same web page or different pages of the same website to provide navigational help to surfer of complex a website. All these edges corresponding to non-informative hyperlinks must be removed from the graph.

After removal of every non-informative edge, every isolated node, i.e. node with in-out degree as zero must be removed from the graph. Now, from the set of nodes and edges of the refined graph, 100 nodes are selected having a minimum of five out-degrees and two, in-degrees. The set of selected 100 nodes is termed as collection C or seed set R_f for ranking of authority and hub resources according to their relevancy level of TOI. There are two reasons for choosing 30 queries each with 100 nodes, first is that any user hardly visits any result after a few pages (Jansen, Spink, Bateman, & Saracevic, 1998). The second reason is the type and amount of in-hand access domain expertise for relevancy checking.

7. Experimental evaluations

To prove the effectiveness of our proposed system model, we had performed extensive evaluations over a dataset and analyzed different aspects of the model, as we will see in the next sections. In this section, we describe the metric used to evaluate our proposed system model. The optimal value of the parameter used as an input to the proposed model is also estimated here. Proposed algorithm is basically variant of SALSA that uses query specific neighborhood graph as HITS does, and it also computes a hub score and an authority score for each node in the neighborhood graph. We performed PAGE RANK, HITS, SALSA and sNorm(p) computations based on Precision@k, MRR, MAP, DCG@100 and NDCG @100. One would expect that sampling more back-links should improve the effectiveness of PAGERANK, HITS and SALSA, since it leads to a closer approximation of the complete neighborhood graph. However, Fig. 4 shows the retrieval performance of PAGERANK, HITS, SALSA and sNorm(p) authority score as a function of the number of sampled back-links.

7.1. Evaluation metric

We have used the Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), Precision@k, Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (NDCG) as a performance measuring statistical metric, taking into consideration of ranking of web pages in order of their relevancy level. On the basis of relevancy, evaluation metrics are divided into 2 categories. First is binary relevancy, in which there are only two levels of relevancy i.e. 0 or 1, where 0 means a web page is not relevant to query and 1 means web page is relevant. The second category is multiple levels of relevance, in which there are more than two relevancy levels, but here in our experiment, four levels of relevancy are taken, i.e. 0, 1, 2 or 3. Where 0 represents a web page is not relevant to query, 3 represent a highly relevant web page, and 1 and 2 represent a web page has relevancy somewhere between 0 and 3. MRR, MAP and Precision@k come under binary relevancy level whereas multiple levels of relevancy include DCG and NDCG. In the experiment, relevancy level value for each web page is determined in the light of the judgment given by different human participants belonging to different domains.

MRR is computed as the multiplicative inverse of the rank of a relevant first response to a query, averaged across 30 queries. Precision@k is the percentage of relevant web pages up to kth rank. The web pages with a rank lower than the kth rank in a ranking hierarchy are discarded in the calculation. MAP is the mean of Average Precision (AP) of each query in a query set i.e. 30 queries. AP for a query is the average of Precision@k calculated for only the ranks at which web page is relevant to the query. MRR, Preccision@k and MAP are defined as:

$$\text{MRR} = \frac{1}{|q|}\sum_{i=1}^{|q|}\frac{1}{r_i} \tag{10}$$

$$\text{Precision@k} = \frac{rel_t}{k} \tag{11}$$

$$\text{AP} = \frac{1}{|rel_N|}\sum\nolimits_{krel_N}\text{Precision@k} \tag{12}$$

$$\text{MAP} = \frac{1}{|q|}\sum_{i=1}^{|q|}\text{AP}_i \tag{13}$$

where $|q|$ is a total number of queries, r_i is the rank position of a first relevant web page in the list of web pages retrieved for the ith query, rel_t is a total number of relevant web pages up to the kth rank and rel_N is the set of relevant web pages for a query. A value of MRR and MAP ranges from 0 to 1. System is having higher MRR and MAP value as compared to other is more preferable.

Jarvelin and Kekalainen (2002) have given the concept of NDCG which now has become a famous metric used by many web search companies to measure the performance of their web page ranking algorithms. The premise on which NDCG is based is that the web page which is highly relevant a w.r.t. the query issued must be penalized if it is ranked lower in the result list returned by a web page ranking algorithm. Lower the ranking of a highly relevant web page less the usefulness of it for a query issuer, as many studies revealed that a search engine user rarely considers anything relevant or they even did not traverse any result beyond the first few results (Jansen et al., 1998). As a penalty, the graded relevance measure of a web page is discounted or decreased logarithmically proportional to the rank of a web page in the result list. The NDCG values range from 0 to 1, with 1 being an indicator of "perfect" ranking process used by a web search engine. The values of DCG are normalized to give NDCG values by dividing a DCG value for a query by Ideal DCG (IDCG) values for that query. The accumulated gain at a particular rank k i.e. DCG@k is defined as:

$$\text{DCG@}k = \sum_{j=1}^{k} \frac{2^{rel(j)} - 1}{log_2(j + 1)} \tag{14}$$

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k} \tag{15}$$

where $rel(j)$ is the rating (3 = highly relevant, 2 = more relevant, 1 = less relevant and 0 = irrelevant) of a web page. *IDCG@k* is the maximum possible DCG in rank position k, a query can have. It is computed by first sorting all the web pages retrieved as a result of the query by a search engine in-order of their respective relevance rating in descending order. After sorting, DCG@k is computed using equation (Eq. (14)) to give IDCG@k.

7.2. Estimation of parameter p-norm(p)

The results obtained by varying over values of p-norm are illustrated in Figs. 2 and 3. The value of p is passed as a parameter to the algorithm. This parameter allows to control the increase and decreases in the scaling contribution of authority and hub weights while calculating authority weight of the corresponding hub weight or vice versa. $p \in [1, \infty]$ as the value of p increases, p-norm will help high authority and hub weights to dominate the game.

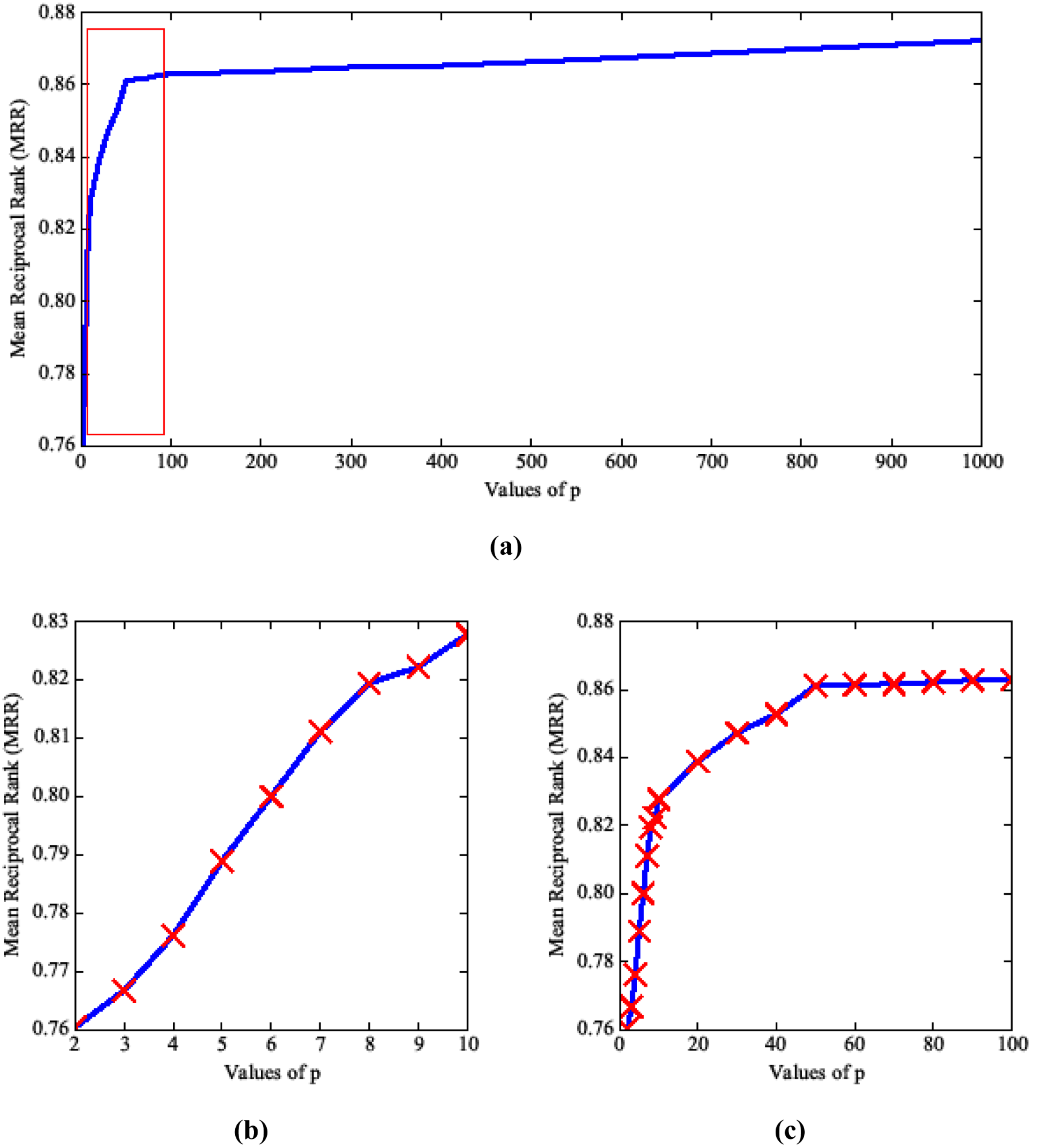
At $p = 1$ alias taxicab norm or Manhattan norm, every component, i.e. authority and hubs will have the same effect in the score calculation and this case is similar to SALSA. At $p = 2$ alias Euclidean norm, larger value components, i.e. high weighted authority and hubs will start being more effective in score calculation for comparison to lower weight. At $p = \infty$ alias infinity norm, only component, i.e. authority and hubs with the largest value will affect the score.

Here, in our experiments value of p is varied from 2 to 1000 as it is very difficult to perform an experiment on the full range with an in-hand accessible resources. But, our selected range has explained the whole game clearly. In Figs. 2(a) and 3(a), we have shown the variation in the values of MRR and MAP at different values of p. Figs. 2(b) and 3(b) and Figs. 2(c) and 3(c) are plotted to show 2–10 and 2–100 values of p clearly as it is not able to see notable variations in values of MRR and MAP in 2–1000 range of p values in Figs. 2(a) and 3(a). Obtained results show that the value of both MRR and MAP are increasing exponentially, for 2–10 values of p as shown in Figs. 2(b) and 3(b). After p equals to 10, MRR increases till p equals to 50 but after that rate of increase becomes very slow as shown in Fig. 2(c). At p equals to 50, the value of MRR is 0.86111 and when p equals to 1000 then it is 0.8722 but at the cost of a very high increase in a number of calculations and computation time. So, p equals to 50 seems to be optimal according to MRR values. The slope of the curve almost becomes flat as shown in Fig. 2(a). But, in the case of MAP, values of y-axis increased till p becomes equal to 60 and after that slope of the curve becomes almost flat due to a decrease in the high rate of increase in MAP value w.r.t. p as shown in Fig. 3(a) and 3(c). At p equals to 60, the value of MAP is 0.794 and at p equals to 1000 then it is 0.80. But, that too at the cost of an increase in a number of calculations and computation time. Therefore, p equals to 60 seems to be optimal according to MAP values.

An ideal value of MRR and MAP is "1" which can be achieved only at very high values of p, but as p increases the level of complexity and computation time of the system will increase with a very high rate. The optimal value of MRR and MAP which can be achieved by the proposed system model is 0.8611 and 0.794 respectively at lower values of p without much increase in complexity and computation time. However, for MRR optimal value is obtained at $p = 50$ and for MAP it is $p = 60$ as shown in Figs. 2(c) and 3(c). So taking in an account of computation time and complexity we will select lower bound to mark $p = 50$ as the most optimal value of p in our experiment. Further, every computation or comparison with the baseline is done at $p = 50$ for our system sNorm(p).

8. Comparison with baselines

Proposed algorithm is basically variant of SALSA that uses query specific neighborhood graph as HITS does, and it also computes a hub score and an authority score for each node in the neighborhood graph as HITS does. However, while HITS uses a methodology called "mutual enforcement" where hubs enforce authorities and vice versa. So, SALSA takes minimum query time because these hub and authority scores calculated by sNorm(p) considered two independent random walks on the neighborhood graph, a hub walk and an authority walk. The set of web pages related to a given query need to be ranked in order of their relevancy level, i.e. more relevant web pages in the top of a ranking hierarchy and less relevant or irrelevant ones at the bottom. The objective here is to analyze how well the proposed system model fulfills user information needed in comparison to state of the art models for ranking of web pages. From the previous section, the optimal value of p, i.e. p =50 and ranking algorithm explained in Section 2 are used to evaluate the proposed system model. Here notable point is that comparisons are performed using a different set of 30 queries which are independent of those that are used to obtain optimal value of p while training the proposed system. The comparison of the proposed system model is performed with PageRank, HITS and SALSA on the basis of Precision@k, MRR, MAP, DCG@100 and NDCG @100.



(a)

(b)

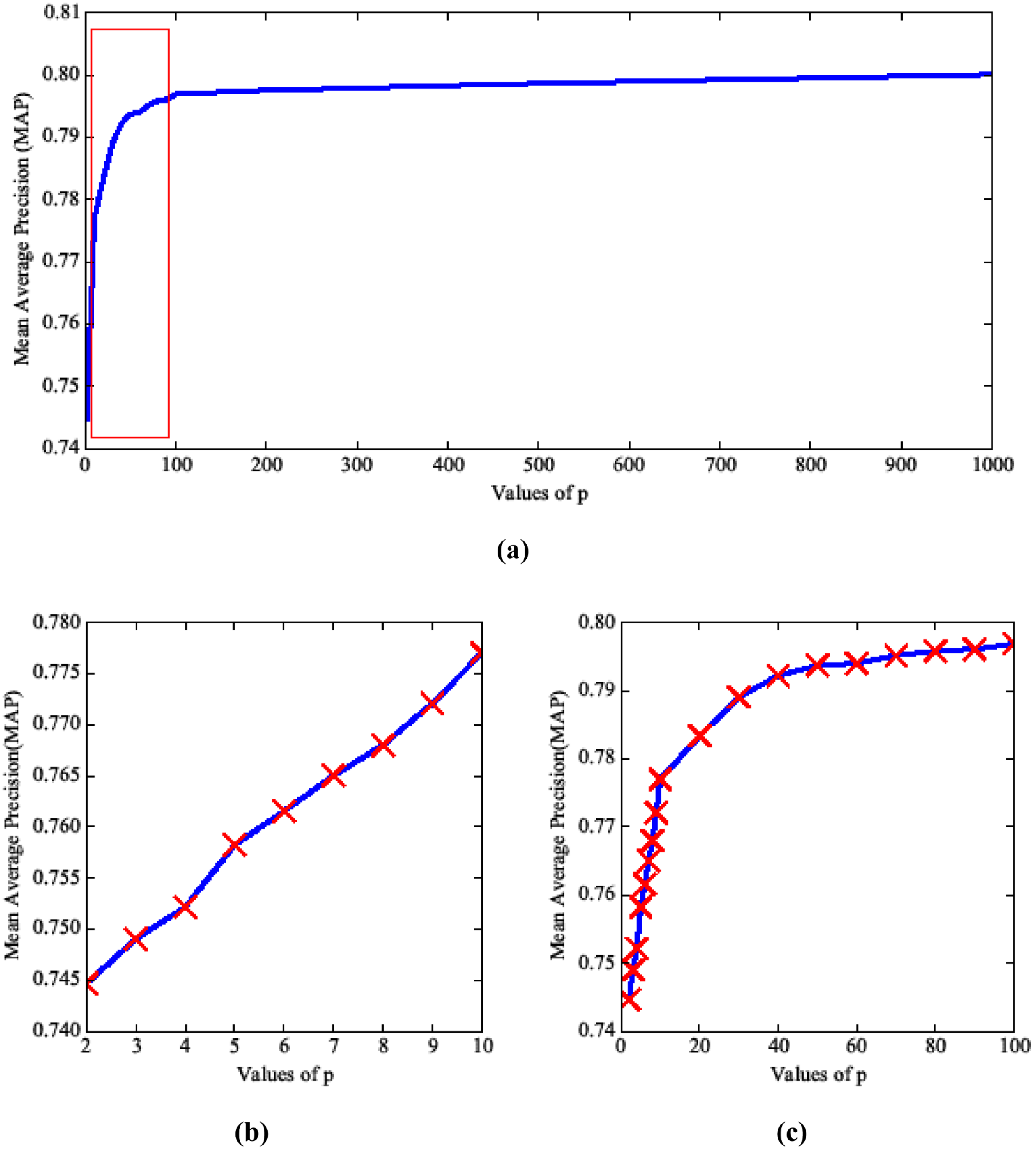
(c)

Fig. 2. Impact of p on MRR (a) scale of p varies in range 2–1000, (b) scale of p varies from 2 to 10, (c) scale of p varies from 2 to 100.

8.1. Results analysis

The proposed system model and selected baseline models stated above have been implemented using Python 2.7.10 Shell. The results have been obtained on a Dell Workstation T5600 – with 2.6 GHz Intel Xenon e5 2650 CPU and 8GB 1600 MHz DDR3 RAM, running windows plkatform. Comparisons of obtained results are illustrated with the help of Tables, bar-plots and line plots.

In Table 2, Precision@k is calculated for each model under consideration at ten different values of k. The value of k is only varied till 80 because a query issuer rarely visits web pages ranked lower in the hierarchy. For each value of k, sNorm(p) proves to be more efficient than baseline models under consideration. For example, at $k = 14$ the precision value of sNorm(p) is 0.928571 which means, 92.85% of web pages among top 14 web pages are relevant if the ranking hierarchy obtained using sNorm(p) as a ranking algorithm



(a)

(b)

(c)

Fig. 3. Impact of *p* on MAP (a) scale of *p* varies in range 2–1000, (b) scale of *p* varies from 2 to 10, (c) scale of *p* varies from 2 to 100.

by a search engine. For PageRank number of relevant web pages among top 14 web pages is 71.428%, HITS is 78.571% and for SALSA it is 84.615%. For the sake of simplicity, Precision@k is only shown for Query1.

In Table 3, the obtained values of MRR and MAP clearly shows that the proposed system model *s*Norm(*p*) is more efficient than its predecessors. The difference in values of *s*Norm(*p*) and its immediate predecessor is nearly 0.0333 units in MRR value and 0.015 units in MAP value. MRR is the mean over the reciprocal of location of 1st relevant web page in the ranking hierarchy over each query into a set of queries. Here, the value of MRR is 0.8666 for *s*Norm(*p*) that means 86.66% of queries among the query set has a relevant web page at rank = 1; that percentage of obtaining relevant web page earlier in hierarchy is 86.66%. For PageRank number of queries with relevant web pages at rank = 1 is 78.33%, HITS is 81.66% and for SALSA it is 83.33%. Similar to precision it also represents the percentage of relevant web pages among top few results returned by search engines, but in MAP, average precision is averaged over

Table 2
Precision@k for Query1.

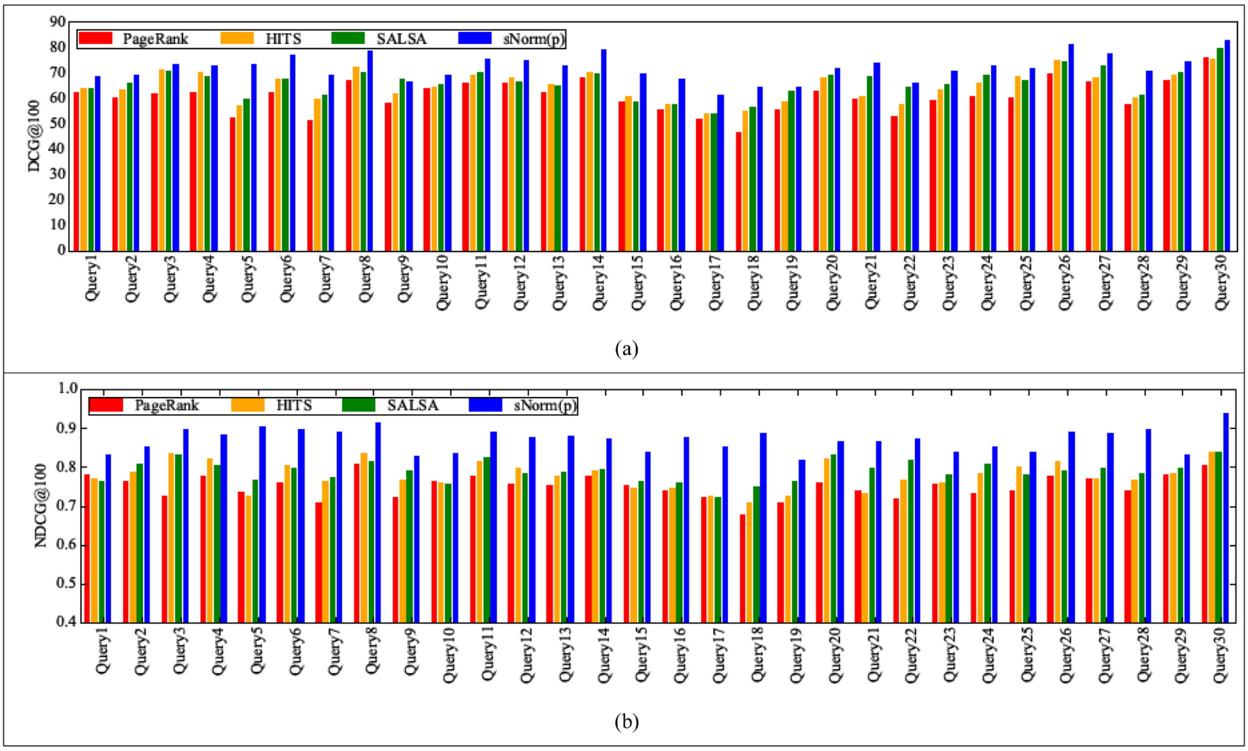
Precision@k	PageRank	HITS	SALSA	*s*Norm(p)
Precision@9	0.888889	1	0.888889	1
Precision@14	0.714286	0.785714	0.846154	0.928571
Precision@22	0.681818	0.772727	0.818182	0.863636
Precision@30	0.633333	0.700000	0.800000	0.833333
Precision@38	0.684211	0.710526	0.842105	0.815789
Precision@44	0.659091	0.727273	0.840909	0.840909
Precision@52	0.692308	0.750000	0.826923	0.846154
Precision@60	0.700000	0.733333	0.816667	0.850000
Precision@74	0.729730	0.743243	0.797297	0.851351
Precision@80	0.737500	0.750000	0.800000	0.837500

Table 3
Comparison of system model based on MRR and MAP.

System models	Mean Reciprocal Rank (MRR)	Mean Average Precision (MAP)
PageRank	0.7833	0.748453
HITS	0.8166	0.778452
SALSA	0.8333	0.780764
*s*Norm(*p*)	0.8666	0.795701

the number of queries in a query set. The sole reason behind this is to compare the performance of different system models used for web page ranking over a set of multiple queries. For some queries, one system model wins the race and for other queries, another model wins the race. So, to assess the overall performance of both models their MAP value is computed. Here, MAP value of proposed *s*Norm(*p*) is more than other system models as shown in Table 3, so on the basis of MAP *s*Norm(*p*) is more efficient. Precision@k, MRR, and MAP are measures based on the binary relevancy of web pages. A system with higher MRR and MAP value is always preferable over models having lower values.

In Fig. 4, bar-plots clearly shows the difference in DCG and NDCG values made by PageRank, HITS, SALSA and proposed *s*Norm(*p*) w.r.t different queries. Both DCG and NDCG for a query are computed at 100th rank, therefore, DCG@100 and NDCG@100 notations



(a)

(b)

Fig. 4. Analysis of (a) DCG@100 (b) NDCG@100 for 30 quires with 100 web pages each.

are used. For each subplot, X-axis indicates the queries in the query set, while Y-axis in Fig. 4(a) represents the DCG@100 values and in Fig. 4(b) it represents NDCG@100 values. Both DCG@100 and NDCG@100 uses the graded level of relevancy of a web page to judge the usefulness of a web page. Four grading levels, i.e. 0, 1, 2 and 3 are used here for the calculation of DCG and NDCG. Where 3 means highly relevant web page, 0 means irrelevant and 1, 2 means something in between 0 and 3. It is a fact that more relevant web pages must appear on top of the ranking hierarchy of web pages returned as a result of an issued query by the search engine. So in any case, if a ranking hierarchy does not adhere to the fact stated above that means a relevant web page is ranked lower in the hierarchy than DCG will penalize ranking hierarchy. As a penalty, graded relevancy of a relevant web page ranked lower in the hierarchy is decreased in logarithmic proportion. More the value of DCG of a query less the penalization of ranking hierarchy returned as a result of that query by a search engine. Here, values of DCG for each and every query in a query set are more than the other state of the art methods as shown in Fig. 4(a). Therefore, the amount of penalization in the case of proposed sNorm(p) is less which depicts chances of ranking a relevant web page at a lower rank is also less in comparison to other system models.

The NDCG for the query is computed by dividing DCG value by the maximum DCG i.e. IDCG value that query can achieve. The purpose of computing in NDGC is that a performance comparison of a search engine using a particular system model for ranking of web pages for one query to another query cannot be only done on the basis of DCG alone. Therefore, the total accumulated gain at a particular rank position k must be normalized across queries in a query set. Now the values of NDCG are between 0 and 1, so the cross query comparison can also be made easily. NDCG is also a very useful measure, if a number of results returned for different queries are not same. In case of NDCG also sNorm(p) also proves to be more efficient than its counterparts as shown in Fig. 4(b). For the mathematical formulation of Precision@k, MRR, MAP, DCG, and NDCG refer to Section 7.1. So, on the basis of Precision@k, MRR, MAP, DCG and NDCG values discussed above the sNorm(p) leads in every case and proved to be more efficient than PageRank, HITS, and SALSA in the ranking of relevant web pages on the top of a ranking hierarchy.

8.2. Statistical analysis

In order to analyze and verify the performance exhibited by proposed sNorm(p) over baseline models as per the results shown in Tables 2 and 3 and Fig. 4, a robustness test is performed. Robustness of a web page ranking algorithm can be stated as the retrieval rate of relevant web pages in a correct ranking hierarchy. The robustness test is commonly known as t-test among statisticians who uses it to determine that the results obtained under two different conditions are statistically significant or not, it is also treated as a significance test. Here, paired 2-sample t-test (one tail) is performed as same set of queries are considered in both cases i.e. sNorm(p) and baseline models. To avoid biasness or matter of chance, total 40 queries are considered here. Here, to perform the analysis using t-test, an assumption that the population have equal variance at a confidence level of 5% i.e., $\alpha = 0.05$ is considered. The research hypothesis considered to prove the significance of results obtained by sNorm(p) over baselines are as follows.

$$H_0\colon\ \mu_s = \mu_b$$

$$H_1\colon\ \mu_s > \mu_b$$

where, μ_s and μ_b are means of results corresponding to various parameters obtained by sNorm(p) and baselines respectively. Here the test is conducted for precision@k, where $k = 80$, NDCG@100 and DCG@100 parameters. For sNorm(p) the value of $p = 50$ is used. The results of t-test are shown in Tables 4–6.

Table 5.

The observation shown in Tables 4–6 indicates that the t-critical values for different parameters corresponding to baselines under consideration are less than the statistical values (t-stats) obtained by t-test. Also the p-value in every case shown in Tables 4–6 is less than the $\alpha = 0.05$. Thus, there is a significant difference in two type of means which support the claim to reject null hypothesis in favor of our alternate hypothesis.

Therefore, we conclude that efforts introduced by the proposed approach for ranking of web pages bring a considerable improvement of search quality. Note that the proposed system is tested on 100 nodes selected from refined graph obtained after preprocessing on the basis of in-out degree threshold, but it can be easily scaled up for nodes greater than 100. Here, each node refers to a web page. For the web search engine users, major implication is – to realize that search results retrieved by the search engine they are using are not properly ordered according to the relevancy of search results w.r.t TOI. Most of the users have never visited, the results beyond the first page of results retrieved by web search engines (Jansen, Spink, Bateman, & Saracevic, 1998). Due to various

Table 4
Results of t-test for precision metric.

	PageRank	HITS	SALSA	sNorm(p)
Average precision @ k	0.39020	0.42286	0.42367	0.62217
Variance	0.0628	0.0635	0.0864	0.0567
Observation	40	40	40	40
Hypothesized mean difference	0	0	0	
Degree of freedom	39	39	39	
t-stat	3.838	3.601	3.217	
p-value one tail	0	0.001	0.003	
t-critical one tail	1.68	1.68	1.68	

Table 5
Results of *t*-test for NDCG metric.

	PageRank	HITS	SALSA	*s*Norm(*p*)
Average NDGC	0.37961	0.41620	0.45712	0.58480
Variance	0.0387	0.0373	0.0650	0.0618
Observation	40	40	40	40
Hypothesized mean difference	0	0	0	
Degree of freedom	39	39	39	
t-stat	3.813	3.417	2.268	
p-value one tail	0	0.001	0.029	
t-critical one tail	1.68	1.68	1.68	

Table 6
Results of *t*-test for DCG metric.

	PageRank	HITS	SALSA	*s*Norm(*p*)
Average DCG	38.460	40.519	42.886	60.4908
Variance	708.411	603.73	695.751	675.090
Observation	40	40	40	40
Hypothesized mean difference	0	0	0	
Degree of freedom	39	39	39	
t-stat	3.409	4.322	3.344	
p-value one tail	0.002	0	0.002	
t-critical one tail	1.68	1.68	1.68	

link structure analysis problems, there may be chances that some relevant web pages ranked in lower in a hierarchy by the search engine. So to tackle this problem improvement in ranking methodology is required. For search engine developers, major implication is to design a new search engine according to rules of proposed *s*Norm(*p*). If it's not possible to design a new search engine, then embed a re-ranking module for pages according to ranking rules of *s*Norm(*p*).

9. Inferences

In this paper, a novel approach for ranking of web pages is presented, i.e. *s*Norm(*p*) on the basis of identification of authoritative web pages and link structure of seed set R_f. A statistical approach is applied on a bipartite graph of R_f with authority resources on one side and hub resources on another side to identify most relevant authoritative web pages. The proposed *s*Norm(*p*) combines the p-Norm from the family of vector norms to SALSA to increase the effect of higher hub weight in a calculation of authority weight and decrease the effect of low hub weight in the calculation of authority weight. This combination will automatically combine link based analysis and content-based evaluation at one place due to the presence of features of SALSA in *s*Norm(*p*). The experimentation is performed on a dataset acquired after pre-processing of the results collected from initial few pages retrieved for a query by the Google search engine. Based on the type and amount of in-hand domain expertise, 30 queries are designed. The extensive evaluation and result analysis are performed using MRR, Precision@k, MAP, DCG, and NDCG as performance measuring metrics. The proposed system model *s*Norm(*p*) is proving to be more efficient in the ranking of web pages w.r.t. their relevancy level in comparison to state of the art methods, which is also confirmed by our experimental evaluations and result analysis. Further, to justify the significance of evaluated results depicting the leading performance of proposed *s*Norm(*p*) over baseline models, has been validated with help of 2-sample paired (one tail) *t*-test.

Even with the interest of the proposed system model, there are still possible improvements that we can embed in our system. We are investigating the aspect of distributed computing, on multiple servers to decrease computational complexity and adapt our system to higher values of *p*. We are also investigating ways to handle problems of web spamming and link spamming aroused due to link farm, which is done to cheat web page ranking algorithms to increase the rank of un-authoritative web pages.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2019.02.004.

References

Borodin, A., Roberts, G. O., Rosenthal, J. S., & Tsaparas, P. (2005). Link analysis ranking: Algorithms, theory, and experiments. *ACM Transactions on Internet Technology, 5*(1), 231–297.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Campos, R., Dias, G., Jorge, A., & Nunes, C. (2016). GTE-Rank: A time-aware search engine to answer time-sensitive queries. *Information Processing & Management, 52*(2), 273–298.

Doslu, M., & Bingol, H. O. (2016). Context sensitive article ranking with citation context analysis. *Scientometrics, 108*(2), 653–671.

Guha, S. K., Kundu, A., & Duttagupta, R. (2015). Introducing link based weightage for web page ranking. *International Journal of Artificial Life Research, 5*(1), 41–55.

Gupta, A., Dixit, A., & Devi, P. (2015). A novel user preference and feedback based page ranking technique. *Proceedings of the IEEE 2nd international conference on computing for sustainable global development (INDIACom)* (pp. 1335–1340). .

Householder, A. S. (2013). *The theory of matrices in numerical analysis.* Courier Corporation.

Jansen, B. J., & Molina, P. R. (2006). The effectiveness of web search engines for retrieving relevant ecommerce links. *Information Processing & Management, 42*(4), 1075–1098.

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the Association for Information Science and Technology, 61*(8), 1517–1534.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum, 32*(1), 5–17.

Jarvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.

Kumar, R. P., Singh, A. K., & Mohan, A. (2016). *Review of link structure based ranking algorithms and hanging pages. Review of link structure based ranking algorithms and hanging pages*420–459.

Langville, A. N., & Meyer, C. D. (2011). *Google's PageRank and beyond: the science of search engine rankings.* Princeton University Press.

Langville, A. N., & Meyer, C. D. (2012). *Who's# 1? The science of rating and ranking.* Princeton University Press.

Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks, 33*, 387–401.

Li, J.-Q., Zhao, Y., & Garcia-Molina, H. (2012). A path-based approach for web page retrieval. *World Wide Web, 15*(3), 257–283.

Liu, J., Kim, C. S., & Creel, C. (2015). Exploring search task difficulty reasons in different task types and user knowledge groups. *Information Processing & Management, 51*(3), 273–285.

Najork, M., Zaragoza, H., & Taylor, M. J. (2007). Hits on the web: How does it compare? *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 471–478). .

Negm, E., AbdelRahman, S., & Bahgat, R. (2017). PREFCA: A portal retrieval engine based on formal concept analysis. *Information Processing & Management, 53*(1), 203–222.

Schott, J. R. (2016). *Matrix analysis for statistics.* John Wiley & Sons.

Shen, Z. L., Huang, T. Z., Carpentieri, B., Gu, X. M., & Wen, C. (2017). An efficient elimination strategy for solving PageRank problems. *Applied Mathematics and Computation, 298*, 111–122.

Singh, M. K., Sharma, A., Rishi, O. P., & Akhtar, Z. (2017). Knowledge extraction through page rank using web-mining techniques for e-business: A review. *Maximizing Business Performance and Efficiency through Intelligent Systems,* 1–36.

Singh, R., & Sharma, D. K. (2013). Enhanced-ratiorank: Enhancing impact of inlinks and outlinks. *Proceedings of the IEEE international conference on information & communication technologies (ICT)* (pp. 287–291). .

Spink, A., Jansen, B. J., Blakely, C., & Koshman, S. (2006). A study of results overlap and uniqueness among major web search engines. *Information Processing & Management, 42*(5), 1379–1391.

Tagarelli, A., Kabbur, S., & Karypis, G. (2015). *Web search based on ranking. Graph-based social media analysis*67–106.

Thalhammer, A., & Rettinger, A. (2016). Pagerank on wikipedia: Towards general importance scores for entities. *International semantic web conference* (pp. 227–240). .

Wilkinson, J. H. (1965). *The algebraic eigenvalue problem.* Oxford University Press.

Yang, W. (2016). An improved HITS algorithm based on analysis of web page links and web content similarity. *Proceedings of the IEEE international conference on cyberworlds (CW)* (pp. 147–150). .

Yang, W., & Zheng, P. (2016). An improved PageRank algorithm based on time feedback and topic similarity. *Proceedings of the 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (pp. 534–537). .

Zin, T. T., Tin, P., Hama, H., & Toriu, T. (2015). A new look into web page ranking systems. *Genetic and Evolutionary Computing,* 343–351.