Analysis of Link Popularity among the Nepal Governments' Websites by using HITS Algorithm

- **Bhim Bhandari**, PhD Scholar Singhaniya University, 9851053681,
bhim_bhandari@hotmail.com
- **Prof. Dr. Shashidhar Ram Joshi**, Institute of Engineering, IOE,
sashi@healthnet.org.np

ABSTRACT

New innovations in science and technology, communication becomes faster and easier. Information dissemination tools become the URLs through internet. Nepal Government has taken major initiatives and has good impact for g-governance. In this context, a close study among the governmental organizations' URLs and their links from one web page to another helps for information dissemination. A web page's link popularity is measured by the number of other web pages that link to it, and links from pages on other web sites. The page popularity is measured by using HITS and PageRank methods. The two main factors ***relevance*** *and* ***reputation*** *are considered as major findings during study.* ***Relevance*** *is a measure of how easy it is for the search engine to tell that your web page is really about the search term that's been used. It improves relevance with on-page search engine optimization, and talks about that extensively on other pages of our site.* ***Reputation*** *is measured by the number of links coming into one web site and the quality of those links. The most common term for that is "****link popularity****".*

Keyword formulation and link definition in the government web sites is random thus status of link popularity is weak. Out of twenty-nine web sites on five websites have implemented the concept of the inlink and outlink i.e. relevancy is better.

Keyword: *Link analysis, Page popularity, HITS, Page Ranking, Weighted graph, In-degree, Out-degree, link popularity, reputation, relevance, on-page search engine, gov.np, hypertext, hyperlink, tag LL - Link Level.*

INTRODUCTION

Link popularity is the study of total number of web sites that link to defined web page[1]. Webpage link popularity depends on linking from within one web site as well as links from pages on other web sites. Some of the more popular search engines like Google and Yahoo have toolbars which display an indicator of the link popularity of a given page [2].

The link structure of web can be viewed as a graph (link graph) in which each vertex is a web page, and each edge is a hyperlink between two pages. The web graph has some interesting properties, such as power law degree sequence and small diameter [3]. A number of stochastic models for the web graph have been proposed for better understand and predict the statistical properties of the Web [4]. The degree of link popularity refers passing of external links among websites that defines how far a web site gives the related information [5]. Researchers are developing algorithms for, mostly within the scope of single websites such as Wikipedia which may one day be unleashed upon the wider Web without sufficient protection against the creation of hyperlinks that exhibit no useful relevance. To make sure that hyperlink-generating algorithms are up to standard. It needs a way of evaluating hyperlinks that is more subtle than simply comparing the topics of the linked documents [6].

HITS Algorithm

Hyperlink-Induced Topic Search (HITS) algorithm was developed by Jon Kleinberg, a Computer Science Professor at Cornell University. This algorithm calculates the ranking of web pages in an offline mode. It has implemented an online ranking algorithm which approximates HITS [12].

Search engines perform their operations in two phases. In the first phase, this algorithm performs a crawl to gather all the web pages and stores these crawled web pages in the file system. The particular format of storing these web pages differs from one search engine to another. However, these are stored in a compressed format and are indexed for faster retrieval. The next phase involves parsing the content of the stored web pages. This step is essential in order to determine the relative ranking for each page. Ranking the web pages is a highly complex process [13]. Some of the factors that make this complex are the following: billions of web pages, intricate connections among these web pages, different formats, different languages, etc. Apart from these, different search technologies have their own pros and cons. This in turn complicates the functioning of a particular search engine. The HITS algorithm generates ranking for web pages after they have been crawled and stored in a local database.

This process is depicted more clearly in the below diagram:

Fetcher process requests batch file of urls through **Queue Server Process**

a http request process sends from the queue server

Fetcher Process

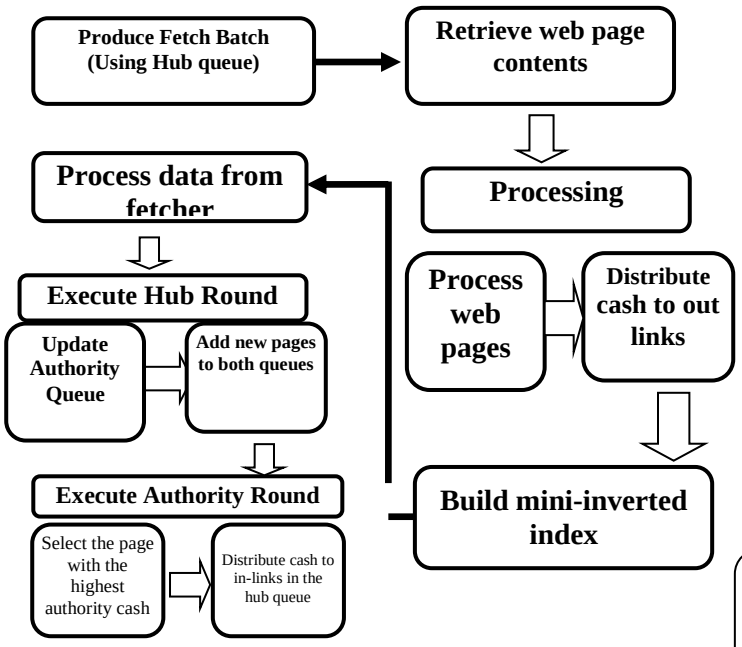

Figure: Overview of Queue Server and Fetcher process

This process is repeated continuously until a sufficient number of web pages are crawled or until terminated by the user. Once the data sent the Fetcher(s) is received by the Queue Server, it is processed and the main index and the priority queues are updated accordingly. Queue Server and the Fetcher can be configured to run on a single host or can be executed on distributed systems. Also, multiple Fetchers can be configured to run simultaneously. This speed up the crawling process and increases the overall efficiency of the system. The advantage of this design is that the script running the Fetcher process needs no modification, as the priority queues are maintained in the Queue Server.

Advantages of HITS

1. HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.
2. The ranking is combined with other information retrieval based rankings.
3. HITS is sensitive to user query (as compared to Page Rank).
4. Important pages are obtained on the basis of calculated authority and hubs value.
5. HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.
6. HITS induces Web graph by finding set of pages with a search on a given query string.
7. Results demonstrate that HITS calculates authority nodes and hubness correctly.

Drawbacks of HITS algorithm

Query Time cost, irrelevant authorities and Hubs, Mutually reinforcing relationships between hosts, Topic Drift, Less Feasibility.

METHODOLOGY

Web page contains so many information and citation. It is tedious task to filter the links of the webpage in terms of its domain information. So it is propose a model that is used to do the above task automatically. The model contains 7 steps and finally gives the link popularity of the web page.

Process Model

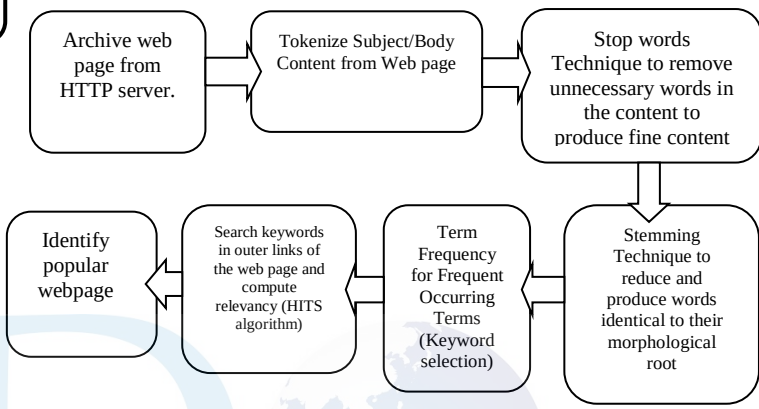

Figure: Process flow in the system model

In the first step of the HITS algorithm the root set (most relevant pages to the query) can be obtained by taking the top n pages returned by a text-based search algorithm. A base set is generated by augmenting the root set with all the web pages that are linked from it and some of the pages that link to it. The web pages in the base set and all hyperlinks among those pages form a focused subgraph. The HITS computation is performed only on this focused subgraph [24]. According to Kleinberg [25], the reason for constructing a base set is to ensure that most (or many) of the strongest authorities are included. The Hub score and Authority score for a node is calculated with the following algorithm [11]:

Pseudo code of HITS algorithm

```
1 Let G be set of pages
2   for each page pg in G do
3       pg.auth = 1
4       pg.hub = 1
5   function Calc_Hubs_Authorities(G)
6    for step from 1 to i do
7        norm = 0
8        for each page pg in G do
9            pg.auth = 0
10           for each page qg in p.inNeighbors
do
11                   pg.auth += qg.hub
```

```
12        norm += square(pg.auth)
13      norm = sqrt(normal)
14      for each page pg in G do
15          pg.auth = pg.auth / normal
```

keyword/webpage	v1	v2	v3	v4	v5	v6
Prevail	1	0	1	1	1	1
Law	1	0	1	1	1	0
Commission	1	1	1	1	1	1
Justice	0	1	1	0	1	1
Prevention	1	0	1	1	1	1
Policy	1	1	1	0	1	1

```
16          norm = 0
17      for each page pg in G do
18          pg.hub = 0
19          for each page rg in pg.outNeighbors
    do
20              pg.hub += rg.auth
21          norm += square(pg.hub)
22      norm = sqrt(normal)
23      for each page pg in G do
24      pg.hub = pg.hub / normal
```

IMPLEMENTTION AND RESULTS

The HITS algorithm is a very popular and effective algorithm to rank documents based on the link information among a set of websites. However, it assigns every link with the same weight which results in topic drift. This paper generalizes the similarity of web pages and proposes a query-induced similarity describing how a webpage is similar to another on a query topic. Then, it also analyzed a new improved weighted hits-based (HITS) algorithm by assigning appropriate weights to link with the similarity and popularity of web pages. Experiment results indicate that the improved HITS algorithm can find more relevant pages than HITS, ARC, SALSA and improve the relevance by 30%-50%.

Tools: The user interface of the thesis was built using jsp with spring framework. The whole system was built in eclipse with jdk 1.6. It has used the open source MySql server as our database.

J2EE: Java two enterprise editions is one of the platforms apart from other two platforms J2SE and J2ME that covers many areas of enterprise and distributed development. The J2EE platform offers a multitier distributed application model, reusable components, a unified security model, flexible transaction control, and web services support through integrated data interchange on Extensible Markup Language (XML)-based open standards and protocols.

Data Samples: The goal of the paper is to rank the Nepal government's web pages on the basis of computing domain relevancy in the outer links of the web pages. There are about fifty governmental web sites and have taken only 29 ministry web sites for the experiment and among them have taken five web sites and their outer links (v1,v2,v3,v4,v5,v6) sample data in the form of adjacent matrices is given as:

Keyword/webpage	v1	v2	v3	v4	v5	v6
Education	1	0	1	1	1	0
Implement	1	1	1	0	1	1
Monitoring	1	0	1	1	0	1
Responsible	0	1	0	0	1	1
Manage	0	1	1	0	1	0
Autonomous	1	0	0	1	1	1

Table: Adjacent Matrix of moe.gov.np web page.

Of

Table: Adjacent Matrix of lawcommission.gov.np web page.

Steps to run Application software

For the detail implementation of the weighted HITS algorithm to compute the link degree among the selected Nepal government web sites, simple application is developed. For the computation, first of all adjacency matrix is computed through the in-degree and out-degree. Each web site is considered as vertex or node. Here, maximum 10 nodes are considered in implementation as below.

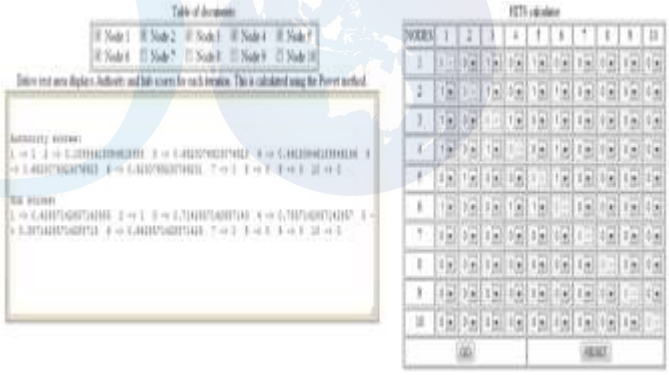

Figure: Authority and hub score of moe

Comparative Study of link levels with associated keywords

		LL_1	LL_2	LL_3	LL_4	LL_5	L
Mohp	Authority	0.66	0.77	0.44	0.22	1.00	0.
	Scores	0.44	0.44	0.33	0.77	0.00	0.
Moe	Authority	0.54	0.71	0.41	0.12	1.00	0.
	Scores	0.39	0.39	0.24	0.77	0.00	1.
Moic	Authority	0.80	0.92	0.80	1.00	0.92	0.
	Scores	1.00	0.95	1.00	0.62	0.95	0.
Mofa	Authority	0.05	1.00	0.44	0.50	1.00	0.
	Scores	1.00	0.95	1.00	0.62	0.95	0.
Law-commission	Authority	0.81	0.63	0.95	0.77	0.81	1.
	Scores	0.86	0.86	0.95	0.77	0.81	1.

Figure: Comparative Study of link with associated keywords

Study of websites with their authority score by using keywords

	Authority				
Keywords	Mohp	moe	moic	mofa	lawcommissio
Link_level_1	0.66	0.54	0.80	0.05	0.81
Link_level_2	0.77	0.71	0.92	1.00	0.63
Link_level_3	0.44	0.41	0.80	0.44	0.95
Link_level_4	0.22	0.12	1.00	0.50	0.77
Link_level_5	1.00	1.00	0.92	0.45	0.81
Link_level_6	0.33	0.30	0.92	0.66	1.00

Figure: Measured Authority Score of web sites.

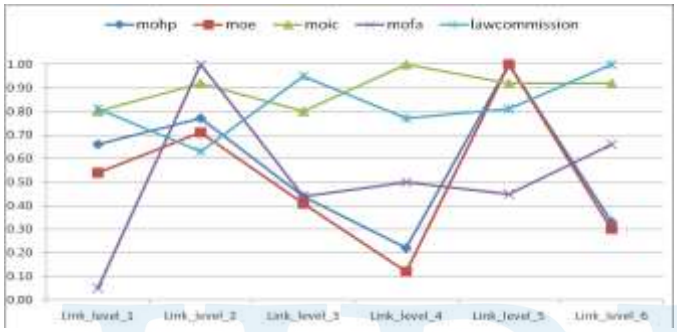

Figure: Measured Authority of different web sites.

From the above line graph analysis, the ministry of information and Communication has the highest authoritative index because all the 6 keywords are ranged within the index value of .80 to 1.00. It means website is well designed with necessary study of the indegree, outdegree, link analysis by ***relevance*** and ***reputation.***

Study of websites with Hub Scores of keywords

	Scores(Hub)				
Keywords	Mohp	moe	Moic	Mofa	Lawcommission
Link_level_1	0.44	0.39	1.00	1.00	0.86
Link_level_2	0.44	0.39	0.95	0.95	0.86
Link_level_3	0.33	0.24	1.00	1.00	0.95
Link_level_4	0.77	0.77	0.62	0.62	0.77
Link_level_5	0.00	0.00	0.95	0.95	0.81
Link_level_6	0.00	1.00	0.95	0.95	1.00

Figure: Measured Hub score of different web sites.

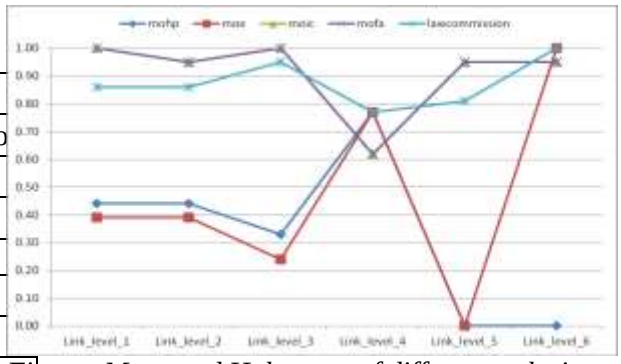

Figure: Measured Hub score of different web sites.

From the above line graph analysis, the ministry of law has the highest score index because all the 6 keywords scores are ranged within the index value of about .70 to 1.00. It means website is well organised from the view of user accessibility. Hence, contents and its links are well managed in ministiry of law and commission**.**

Study of websites with their Authority and hub Scores generated by HITS algorithm

	mohp	Moe	Moic	mofa	lawcommission
Hub	0.30	0.43	0.84	0.84	0.81
Autority	0.35	0.50	0.97	0.97	0.93

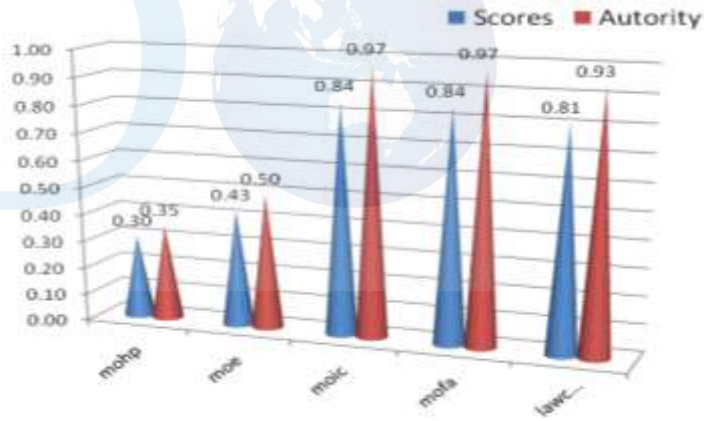

Figure: Histogram measuring authorith and hub score of diffent web sites

From the overall index of authoritative scores analysis of existing status of 5 most common Nepal Government websites, ministry of information and communication and ministry of foreign affairs observed good status on link popularity. The ministry of law also seems to be well organised but needs some improvements. But majority of other ministries websites need more structural, security and link level study and implementation to proper dissipation of the electronic information in the globally competent world.

CONCLUSIONS

This study included twenty nine Nepal governments' ministry web sites for the experiment but only five web sites have their relevancy value well. Among five web sites the best relevancy score has been found in web site Ministry of Information and Communication

(www.moic.gov.np) and other have slight variation in relevancy score. The relevancy score of the web sites shows that only few governmental web sites have their domain related information i.e. the outer links of the web site give the domain related information.

From the overall index of authoritative scores analysis of existing status of 5 most common Nepal Government websites, ministry of information and communiation and ministry of foreign affairs observed good status on link popularity. The ministry of law also seems to be well organised but needs some improvements. But majority of other ministries websites need more structural, security and link level study and implementation to proper dissipation of the electronic information in the globally competent world.

REFERENCES

[1] Erik-Jan van Barren, 2013, Master thesis "Wiki Bench: A, Wikipedia based web application benchmark", 2015.

[2] Studying blog features over Link popularity by Jose Luis Devezas, 2014.

[3] Graham Klyne and Jeremy J. Carroll, editors. Resource Description Framework: Concepts and Abstract Syntax.W3CRecommendation, February 2004.

[4] Wong W., Liu W. & Bennamoun M., 2014. Acquiring Semantic Relations using the Web for Constructing Lightweight Ontologies. In: 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD); Bangkok, Thailand.

[5] Comparative Study of Web 1.0, Web 2.0 and Web 3.0, Umesha Naik D Shivalingaiah Technivision Knowledge Base **Search Engines** A Brief Overview of How They Work In Everyday English! January 1, 2009 Prepared by: Kevin MacDonald

[6] A Novel Architecture of Ontology-based Semantic Web Crawler, Ram Kumar Rana IIMT Institute of Engg. & Technology, Meerut, India, Nidhi Tyagi Shobhit University, Meerut, India

[7] Ranking Techniques for Social Networking Sites based on Popularity, Mercy Paul Selvan et al / Indian Journal of Computer Science and Engineering (IJCSE).

[8] "Survey on Web Page Ranking Algorithms", Mercy Paul Selvan, A .Chandra Sekar, A.Priya Dharshin *International Journal of Computer Applications (0975 – 8887) Volume 41– No.19, March 2012*

[9] PageRank explained *or "Everything 've always wanted to know about PageRank"* Written and theorised by Chris Ridings.

[10] Association Rule Mining based on Ontological Relational Weights, N. Radhika, K.Vidya, Department of Computer Science and Engineering, Aurora's Technological and Research Institute, India.

[11] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins, *The Web as a graph: measurements, models and methods*, Proc. Fifth Ann. Int. Computing and Combinatorics Conf., Springer Verlag Lecture Notes in Computer Science

[12] "Finding Authorities and Hubs From Link Structures on the World Wide Web" by-Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas

[13] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[14] "A Syntactic Classification based Web Page Ranking Algorithm", Debajyoti Mukhopadhyay, Pradipta Biswas, Young-Chon Kim

[15] "Modeling and Optimizing Hypertextual Search Engines" Based on the Research of Larry Page and Sergey Brin, Yunfei Zhao Department of Computer Science, University of Vermont Slides from Spring 2009 Presenter: Michael Karpeles

[16] Google and the Page Rank Algorithm, slides by Székely Endre 2007. 01. 18.

[17] Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search Taher H. Haveliwala Stanford University taherh@cs.stanford.edu

[18] The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank Matthew Richardson Pedro Domingos Department of Computer Science and Engineering University of Washington Box 352350 Seattle, WA 98195-2350, USA *{mattr, pedrod}@cs.washington.edu*

[19] Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization Rada Mihalcea Department of Computer Science University of North Texas rada@cs.unt.edu

[20] Hyperlink Analysis: Techniques and Applications Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and PangNing Tan Department of Computer Science, University of Minnesota, Minneapolis, MN, USA

[21] Comparative Analysis of Pagerank and HITS Algorithms**,** Nidhi Grover, Ritika Wason, MCA Scholar, Institute of Information Technology and Management.