International Journal of Latest Research in Science and Technology **ISSN (Online):2278-5299**
Volume 4, Issue 1: Page No.70-74, January-February 2015
https://www.mnkpublication.com/journal/ijlrst/index.php

i-SEARCH: DOCUMENT SEARCHING USING PAGERANK ALGORITHM

[1]Renalyn C. Antonio, [2]Lorena W. Rabago, [3]Bartolome T. Tanguilig III
[1]Technological Institute of the Philippines, Quezon City, Philippines
[2] Technological Institute of the Philippines, Quezon City, Philippines
[3] Technological Institute of the Philippines, Quezon City, Philippines

Abstract- The study focuses on optimizing document searching using PageRank algorithm for White Hat Search Engine (SEO) technique for digitized master's theses and doctoral dissertations of the Technological Institute of the Philippines-Quezon City Graduate Programs. This project will be beneficial to the readers, authors, and researchers for their research related needs like review of related literature and among others. Several techniques and technologies were used for the development of the project. The SDLC waterfall method served as the model in developing the system. PHP:Hypertext Preprocessor was used as the scripting language and MySQL as the database management system. White Hat SEO technique and strategies which target a human audience opposed to a search engine was also used in the project. The system was evaluated on its acceptability in terms of usability, functionality, reliability using ISO 9126. The respondents rated the system highly acceptable in terms of optimizing their search activities. This is a clear indication that optimizing document searching for any data or documents repository should be optimized for better customer service.

Keywords - Page Rank Algorithm; Search Engine Optimization; White Hat SEO techniques; Repository; Document Management System

I. INTRODUCTION

This paper focuses on the development of Document Management System (DMS) for graduate programs electronic theses and doctoral dissertations. The project implemented White Hat SEO technique utilizing Page Rank algorithm in accessing relevant data from DMS of theses and dissertation. In the previous years, searching of relevant information from the compilation of theses and dissertations of the graduate programs was slow and tedious. At present, students produce a bound and CD copies of their output which are available only within the campus; access to such scholarly outputs is limited, not widely accessible and disseminated.

Now, with the available tools to optimize the searching process, access is more convenient. Higher Education Institutions' environment is changing, thus the need for a DMS is inevitable. This endeavor is beneficial to the viewers, authors, educators and researchers worldwide to further enhance their knowledge in coming up with research output.

This study aims to come up with a model or electronic system that would provide fast access and long-term archiving / retention of masters theses and dissertations of TIP graduate programs' students making it available online, theses and dissertations widen its accessibility and usefulness for researchers worldwide.

Theses and dissertations (TDs) were preserved for future use; conserve paper and storage spaces; help avoid duplication of topics and area of research; easy monitoring of theses and dissertations; systematic and organized TDs; easy retrieval, enhanced security and secure back-ups. The authors utilized White Hat SEO technique in developing the system. SEO helps in positioning the website properly for easy referencing when researchers need the site. The authors also optimized document searching using PageRank algorithm.

II. REVIEW OF RELATED LITERATURE

There are two major pillars of search engine optimization:

A. On-Page SEO

It includes providing good content, good keywords selection, putting keywords on correct places, giving appropriate title to every page [6] or site structure. This ensures that everything is done on the actual webpage to improve the rank.

B. Off-Page SEO

It includes link building, increasing link popularity by submitting open directories, search engines, link exchange [6] and social media. This is doing everything off the page to improve the rank.

Search Engine Optimization is one field to attain results; we need to get updated with the algorithms if the search engines change from time to time. This means that the strategy that is followed some years back are now out dated and may look as spam to the crawlers. In this paper, recommendations and building links with different anchor texts from all link like article submissions, forum links, web 2.0 sites, blog commenting etc for diversity in link structure were stressed[11]. PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google.

Publication History
Manuscript Received : 17 February 2015
Manuscript Accepted : 23 February 2015
Revision Received : 23 February 2015
Manuscript Published : 28 February 2015

PageRank is a way of measuring the importance of website pages [13]. According to Google: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites [13].

III. SEARCH ENGINE OPTIMIZATION

In order to meet the goal of search engines to locate the most relevant content, the authors utilized a search engine optimization (SEO) technique in designing the system. SEO is the process of improving the visibility of a website or a webpage in a search engines results page [8]. SEO is fundamental as it creates great and seamless user experience. SEO helps in positioning the website properly to be found at the easiest way when people need the site.

Fig. 1 Search Engine Optimization Process

IV. SEARCH ENGINE BASICS

Search engines perform three key processes to deliver results: crawling, indexing, and ranking.

A. *Crawling*

It is the process by which search engines discover updated content on the web, such as new sites or pages, changes to existing sites, and dead links [2]. To do this, a search engine uses a program that can be referred to as a 'crawler', 'bot' or 'spider' which follows an algorithmic process to determine which sites to crawl and how often. Web crawlers – also known as robots, spiders, worms, walkers, and wanderers – are almost as old as the Web itself [5]. Process of fetching all the web pages linked to a website. This task is performed by a software called a crawler or a spider [6].

B. *Indexing*

After a page is crawled, the next step is to index the content [1]. This indexed content is then stored, with the information then organized and interpreted by the search engine's algorithm to measure its importance compared to similar pages [2]. Process of creating index for all the fetched web pages and keeping them into a giant database from where it can later be retrieved. Essentially, the process of indexing is identifying the words and expressions that best describe the page and assigning the page to particular keywords [6].

C. *Ranking*

Once a keyword is entered into a search box, search engines will check for pages within their index that are a closest match; a score will be assigned to these pages based on an algorithm consisting of hundreds of different ranking signals [2].

V. METHODOLOGY

A. *System Development Life Cycle (SDLC)*

SDLC (Systems Development Life-Cycle) is used in information systems, systems engineering, and software engineering as a process of creating new or altering existing systems [14]. The SDLC can be thought of as a concept that lies beneath a number of software development methodologies currently employed throughout industry. From these, the framework to create, plan, and control an information system flows which is also known as the software development process [14]. The SDLC waterfall method served as the model in developing the system. Waterfall Model was used because it is a conceptual model that serves as the guideline in developing the researchers' system. It is a linear sequential life cycle that each phase must be completed in its entirety before the next phase begins.

B. *White Hat Search Engine Optimization Technique*

Among the several search engine optimization techniques; the authors utilized White Hat SEO Technique during the development of the systems. This refers to the use of techniques and strategies that target a human audience as opposed to a search engine [9]. Techniques that are typically used in white hat SEO include using keywords, and keyword analysis, doing research, rewriting meta tags in order for them to be more relevant, backlinking, link building as well as writing content for human readers. Those who use white hat SEO expect to make a long-term investment on their website, as the results last a long time. White hat SEO refers to the use of good practice methods to achieve high search engine rankings. They comply with search engine guidelines [10].

C. *Page Rank Algorithm*

The PageRank algorithm was invented by Page and Brin around 1998 and used in the prototype of Google's search engine. The objective is to estimate the popularity, or the importance, of a webpage, based on the interconnection of the web [3]. The original PageRank algorithm was described by Lawrence Page and Sergey Brin in several publications [5]. PageRank is a measure of the popularity of a webpage as determined by the hyperlinks from other pages leading to it, as well as the popularity of those linking pages themselves [7]. It is given by:

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

Where:

PR(A) is the PageRank of page A

PR(Ti) is the PageRank of pages Ti which link to page A

C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1

A simple way of representing the formula is, (d=0.85) *Page Rank (PR) = 0.15 + 0.85 * (a share of the Page Rank of every page that links to it).*The amount of Page Rank that a page has to vote will be its own value * 0.85. This value is shared equally among all the pages that it links to [5].

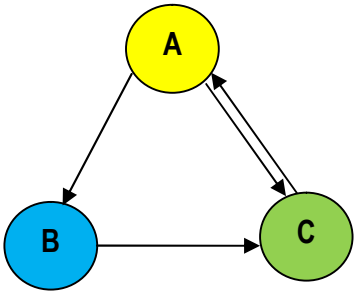

Fig. 2 Page Ranking

Based on the above figure, this study regarded a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85.

The computation of PageRank values is approximative and iterative because the size of the web pages varies. Where each page is assigned initial starting value and Page rank of all pages are calculated using the algorithm.

VI. RESULTS AND DISCUSSION

Results and discussion of data are presented in this section. It is indeed advantageous to utilize PageRank algorithm because it is one of the factors that determines a page's ranking in the search results. The ranking of a web page depends on rank of other web pages pointing to it. Even though there are fewer links to a certain page compared to other pages, if it comes from an important page and rank higher than other pages, then the search was already optimized.

Below is the illustration of the implementation of PageRank algorithm using two different webpages. Each page has one outgoing link so that means each back link is equal to one (1). The illustration shows that Page Rank of A depends on Page Rank value of B and vice versa. Thus, it is obvious that they are both dependent on the importance of each other.

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn)$$

$$\begin{aligned} PR(A) &= (1- d) + d\,(PR(B) / C(B) \\ &= (1 – d) + d(PR(B)/1) \\ &= (1-.85) + .85\,(1/1) \\ &= .15 + .85 \\ &= 1 \end{aligned}$$

$$\begin{aligned} PR(B) &= (1- d) + d\,(PR(A) / C(A) \\ &= (1 – d) + d(PR(A)/1 \\ &= (1-.85) + .85\,(1/1) \\ &= .15 + .85 \\ &= 1 \end{aligned}$$

Below is another illustration using PageRank algorithm for three different web pages with several iterations using the damping factor of 0.85.

$$PR(A) = 0.15 + 0.85\,PR(C)$$
$$PR(B) = 0.15 + 0.85\,(PR(A) / 2)$$
$$PR(C) = 0.15 + 0.85\,(PR(A) / 2 + PR(B))$$

TABLE 1 - THE ITERATIVE COMPUTATION OF PAGERANK

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.575	1.064
2	1.0541875	0.598029688	1.10635492
3	1.0904017	0.613420716	1.13482832
4	1.1146041	0.623706732	1.15385745
5	1.1307788	0.630581005	1.16657486
6	1.1415886	0.635175168	1.17507406
7	1.148813	0.638245505	1.18075418
8	1.1536411	0.640297449	1.18455028
9	1.1568677	0.641668789	1.18708726
10	1.1590242	0.642585272	1.18878275
11	1.1604653	0.64319777	1.18991587
12	1.1614285	0.64360711	1.19067315
13	1.1620722	0.643880676	1.19117925
14	1.1625024	0.644063505	1.19151748
15	1.1627899	0.644185691	1.19174353
17	1.1631104	0.644321923	1.19199556
18	1.1631962	0.644358395	1.19206303
19	1.1632536	0.64438277	1.19210812
20	1.1632919	0.64439906	1.19213826
21	1.1633175	0.644409947	1.19215840
22	1.1633346	0.644417223	1.19217186

23	1.1633461	0.644422085	1.19218086
24	1.1633537	0.644425335	1.19218687
25	1.1633588	0.644427507	1.19219089
26	1.1633623	0.644428958	1.19219357
27	1.1633645	0.644429928	1.19219537
28	1.1633661	0.644430576	1.19219657
29	1.1633671	0.644431009	1.19219737
30	1.1633678	0.644431299	1.19219790
31	1.1633682	0.644431493	1.19219826
32	1.1633685	0.644431622	1.19219850
33	1.1633687	0.644431708	1.19219866
34	1.1633689	0.644431766	1.19219877
35	1.163369	0.644431805	1.19219884
36	1.163369	0.64443183	1.19219889
37	1.1633691	0.644431848	1.19219892
38	1.1633691	0.644431859	1.19219894
39	1.1633691	0.644431867	1.19219895
40	1.1633691	0.644431872	1.19219896
41	1.1633691	0.644431875	1.19219897
42	1.1633691	0.644431878	1.19219897
43	1.1633691	0.644431879	1.19219898
44	1.1633691	0.64443188	1.19219898
45	1.1633691	0.644431881	1.19219898
46	1.1633691	0.644431881	1.19219898
47	1.1633691	0.644431882	1.19219898
48	1.1633691	0.644431882	1.19219898
49	1.1633691	0.644431882	1.19219898
50	1.1633691	0.644431882	1.19219898
51	1.1633691	0.644431882	1.19219898
52	1.1633691	0.644431882	1.19219898

Table 1 shows that in searching using three webpages, the authors came up with 52 iterations to optimize the search process. Number of iterations is dependent on the number of webpages in the searching process.

The sample screenshots below show the result of implementing PageRank algorithm.

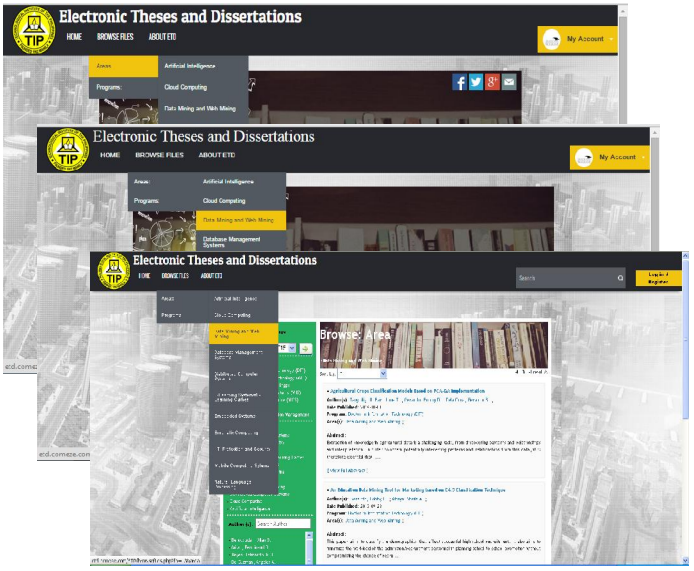

Fig. 3 Search Result per Area

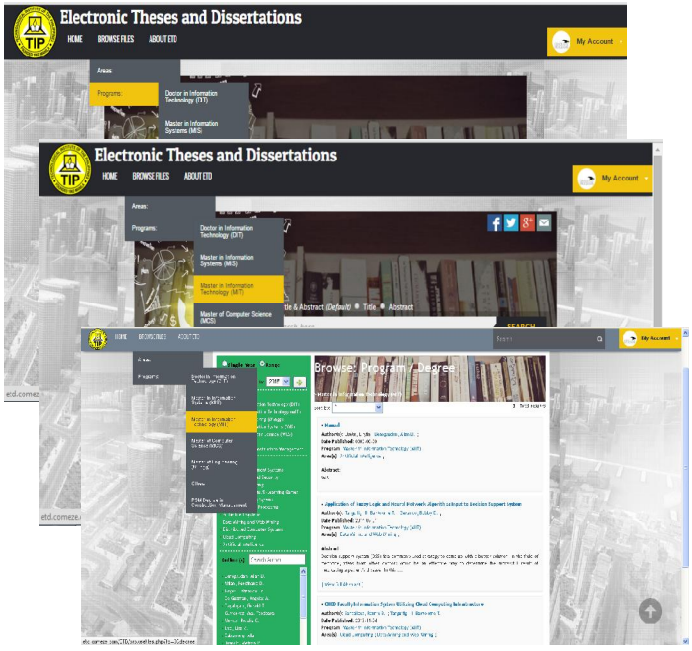

Fig. 4 Search Result per Program

TABLE 2 – SURVEY RESULT

No.	Characteristics	Mean	Interpretation
1	Usability	4.04	Very Usable
2	Functionality	4.05	Very Functional
3	Reliability	4.00	Very Reliable
4	Efficiency	4.20	Very Efficient

Average Mean	**4.07**	Highly Acceptable

Table 2 shows the result of the evaluation from the respondents in terms of the usability of the system. All the given criteria were rated very usable – 4.04, which indicates that the system is indeed useful and easy to understand in terms of accessing relevant information on any research area. It was also noted that the system has a user-friendly interface which makes it easy for the user to navigate. Therefore, the system is indeed highly usable in terms of understandability, learnability and user-friendliness. The respondents also rated the system very functional – 4.05, which indicates that it provides accurate search results and also addresses the need for different searching options. On the other hand, reliability was rated 4.0, which means that fault tolerance or the ability of the system to maintain a specified level of performance in case of software faults is available. Likewise, the efficiency of the system was rated 4.20, which clearly indicates that search results provided by the system are accurate and efficient in terms of its completeness, correctness and timeliness.

In general, the system got an average mean of 4.07 which shows that it is highly acceptable in terms of its functionality, usability, reliability and efficiency.

VII. CONCLUSION

The authors were able to digitize the Technological Institute of the Philippines graduate programs theses and dissertation. White Hat SEO technique using PageRank algorithm was successfully implemented to optimize the searching process. The respondents rated the system as highly acceptable in terms of its usability, functionality, reliability and efficiency.

VIII. RECOMMENDATION

For future studies, it is recommended that other SEO techniques and searching algorithms be used to further enhance the searching process.

ACKNOWLEDGMENT

This endeavor is made possible through the help and support of all the contributors including superiors, mentors, professors, students, family and friends.

REFERENCES

[1] N. Kaur, and J.Kaur, "Development of Ranking Algorithm for Search Engine Optimization," International Journal of Engineering Research & Technology, India, ISSN: 2278-0181,Vol. 3 Issue 4, April 2014.

[2] A. Shore, Search Engine Basics: Crawling, Indexing and Ranking. Totally Communications7. October 2013

[3] K. Shum, "Notes on PageRank Algorithm," ENGG2012B Advanced Engineering Mathematics, April 3, 2013

[4] A. Heydon and M. Najork, Mercator: A scalable, extensible Web crawler, Compaq Systems Research Center, 130 Lytton Avenue, Palo Alto, CA 94301, USA, pp.219 World Wide Web 2 (1999) 219–229

[5] The PageRank Algorithm

http://pr.efactory.de/e-pagerank-algorithm.shtml

[6] Search Engine Optimization – tutorialspoint, simply easy learning

http://www.tutorialspoint.com/seo/seo_tutorial.pdf

[7] J.B. Killoran, Tutorial - How to Use Search Engine Optimization Techniques to Increase Website Visibility, IEEE Transactions on Professional Communication, Vol. 56, No. 1, pp. 53-54, March 2013

[8] Michael David. (May 1, 2013) Search Engine Optimization 2013 – Internetrix Research

[9] http://www.wordstream.com/white-hat-seo

[10] http://www.diffen.com/difference/Black_Hat_SEO_vs_White_Hat_SEO

[11] A New Methodology for Search Engine Optimization with out getting Sandboxed

[12] J.Prethi Sagana Poongkode1, V.Nirosha. A Study on various Search Engine Optimization Techniques. Nov .2014

[13] http://en.wikipedia.org/wiki/PageRank

[14] Waterfall Model Advantages, Examples,Phases andmore about software development. http://www.waterfall-model.com/sdlc/