International Journal of Computer and Information Technology (ISSN: 2279 – 0764)
Volume 04 – Issue 01, January 2015

URL Ordering based Performance Evaluation of Web Crawler

Mohd Adil Siddiqui
Department of Computer Science & Engineering
Integral University
Lucknow, India
Email: mdadilsiddiqui [AT] gmail.com

Sudheer Kumar Singh
Department of Computer Science & Engineering
Integral University
Lucknow, India

Abstract— **There are billions of Web pages on World Wide Web which can be accessed via internet. All of us rely on usage of internet for source of information. This source of information is available on web in various forms such as Websites, databases, images, sound, videos and many more. The search results given by search engine are classified on basis of many techniques such as keyword matches, link analysis, or many other techniques. Search engines provide information gathered from their own indexed databases. These indexed databases contain downloaded information from web pages. Whenever a query is provided by user, the information is fetched from these indexed pages. The Web Crawler is used to download and store web pages. Web crawler of these search engines is expert in crawling various Web pages to gather huge source of information. Web Crawler is developed which orders URLs on the basis of their content similarity to a query and structural similarity. Results are provided over five parameters: Top URLs, Precision, Content, Structural and Total Similarity for a keyword.**

Keywords: Web Crawler; URL Ordering; Web Pages

I. INTRODUCTION

As World Wide Web has grown in leaps and bounds, search engines have become an essential tool. Search engines occupy an important role in providing relevant results by searching vast information on World Wide Web [1]. Searching for relevant information on web is not an easy task. There are many different strategies which work on extracting relevant information of URLs, of these there are three ways in which web data can be mined: content which includes text, multimedia etc [2]; usage which includes server logs to access usage pattern [3]; structure which includes analyzing information from link structure of web [4].

There are different information retrieval techniques such as Boolean model, vector space model and statistical model, probabilistic model, etc [2]. Each information retrieval model represents documents and queries differently but they all treat documents and queries as a bag of words. There are various link analysis algorithms which are well known for example PageRank [5], Hubs and Authority [6], etc. Most popular way of site ranking is link analysis. In PageRank [5], link analysis is done by analyzing both in-links and out-links of the pages. Hubs and authorities [6] also known as hyperlink induced topic search (HITS), searches for relevant page which is referenced to by many pages. Semantic relevance should be considered in ranking URL [7]. Most users first click the URL which is more relevant to their query.

Web crawler (also known as a Web spider or Web robot) is a program or automated script which browses the World Wide Web in an automated and methodical manner [8]. The crawling process increases web traffic to a large extent [9]. In order to minimize the network traffic, the web administrator may implement robot exclusion protocol on their websites. The developed web crawler shown in Figure 1 is multi-threaded which downloads a URL from World Wide Web and stores it in repository. It adheres to robot exclusion protocol. It extracts the list of URL from the downloaded web page and adds them to the URL Queue. The Duplicate URL Elimination eliminates any repeated URLs in the URL Queue. Site ordering module then ranks the URL according to structural and content similarity as implemented ordering algorithm. Indexer updated ranking in repository. Crawl Scheduler then selects the new URL to be crawled from the ordered URLs. The user gets the ordering results fetched from the repository.

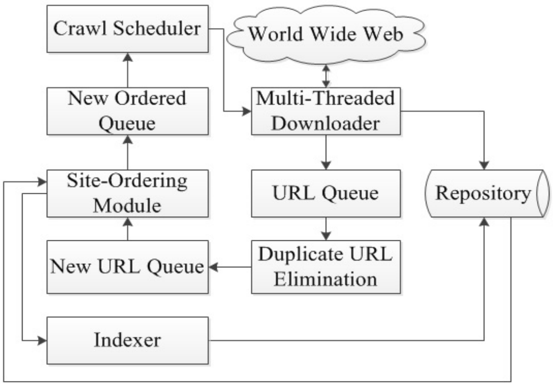


Figure 1. Architecture of Web Crawler

II. PROPOSED ALGORITHM

The proposed algorithm is implemented in site ordering module and uses similarity based approach to rank URLs.

Step 1. Input a URL

An URL is a link to a website. Therefore the URL of website to be crawled is entered by the user to the crawler.

Step 2. Input a Keyword

The user provides a keyword which acts as a query to calculate the content similarity between the pages.

Step 3. Crawl the site

The website whose URL was entered by the user is crawled by crawler to find all the URLs attached to it.

Step 4. Extract the URL in site

All the URLs from the websites are extracted and stored. The content is also extracted and stored. The number of URLs will not be more than the crawl limit set by the user.

Step 5. Calculate Content Similarity

By using TF-IDF [10], calculation of content similarity is done.

a) Term Frequency Scheme (TF): In TF scheme, the weight of a term t_i in a page d_j is the number of times that t_i appears in page d_j. It is denoted by

$$tf_{ij} = \frac{f_{ij}}{\max(f_{1j}, f_{2j}, \ldots, f_{|v|j})} \qquad (1)$$

b) TF-IDF Scheme (TF-IDF): In Inverse document frequency (IDF) is defined as: Total number of pages in web database is denoted by N and the number of pages in which term t_i appears atleast once is denoted by df_i

$$idf_i = \log\frac{N}{df_i} \qquad (2)$$

The formula proposed by Salton and Buckley [11] to calculate TF-IDF weight of each term.

$$w_{ij} = \left\{0.5 + \frac{0.5 * f_{ij}}{\max(f_{1j}, f_{2j}, \ldots, f_{|v|j})}\right\} * \log\frac{N}{df_i} \qquad (3)$$

As it can be seen if term t_i is an important term which appears in every page then $N = df_i$, and $w_{ij} = 0$, which means term t_i weight-age cannot be calculated. Therefore the formula is improved to:

$$w_{ij} = \left\{0.5 + \frac{0.5 * f_{ij}}{\max(f_{1j}, f_{2j}, \ldots, f_{|v|j})}\right\} * \log\frac{N+1}{df_i} \qquad (4)$$

Now in equation, we can see on right-hand part that even if term t_i appears in each document ($\log\frac{N+1}{df_i}$) guarantees that $w_{ij} \neq 0$.

For a query Q, the weight of that query on a page x is denoted as w_{ip_x}. The Content Similarity measure to compute the similarity between pages p_x and p_y can be calculated using:

$$S(P_x, P_y) = \frac{\Sigma w_{ip_x} * \Sigma w_{ip_y}}{(\Sigma w_{ip_x})^2 + (\Sigma w_{ip_y})^2 - (\Sigma w_{ip_x} * \Sigma w_{ip_y})} \qquad (5)$$

Step 6. Calculate Structural Similarity

By using SimRank, structural similarity is calculated.

a) Simrank: SimRank [12] effectively measures similarity by link structure analysis by stating "two objects are similar if they are related to similar objects" [10]. SimRank algorithm analyses the (logical) graphs derived from data sets to compute similarity scores based on the structural context between nodes (objects). The basic concept behind algorithm is that, objects x and y are similar if they are related to objects a and b, respectively, and a and b are themselves similar. The similarity between object x and y is given by s (x, y) ∈ [0, 1]. If x = y then sim (x, y) is defined to be 1. Otherwise

$$sim(x, y) = \frac{c}{|I(x)||I(y)|} \sum_{i=1}^{|I(x)|} \sum_{j=1}^{|I(y)|} s(I_i(x), I_j(y)) \qquad (6)$$

where c is a constant between 0 and 1.

Step 7. Obtain final similarity score

The final similarity score is calculated by:

k_1 * (result of step 5) + k_2 * (result of step 6) (7)

where k_1, k_2 are constants [13]. The value of constants is taken as $k_1 = 0.7$ and $k_2 = 0.3$. As the ordering score gets computed, simultaneously the indexer updates it in the repository. The final score computed is shown as total similarity score.

Step 8. Rank URL according to similarity score

Final rank of all URL along with their respective content, structural and total similarity score is shown as an output.

III. Implementation and Result Analysis

To measure and compare performance of web crawler, web links of four International Journals are taken, and are compared over parameters. The web pages of four international journals were taken up to search a keyword 'journal`. The web pages will be crawled to get appropriate results on basis of similarity measurement for the keyword. The list of international journals is shown in Table I as follows:

TABLE I. List of International Journals

	Name of International Journal	URL of International Journal
1	International Journal of Computer and Information Technology (IJCIT)	http://www.ijcit.com
2	The Science and Information (SAI) Organization	http://www.thesai.org
3	International Journal of Engineering Research (IJER)	http://www.ijer.in
4	International Journal of Soft Computing and Engineering (IJCSE)	http://www.ijsce.org

Parameters taken for result analysis are:

- **Top URLs:** URL list after the crawling process.
- **Crawling time:** Time required crawling and extracting the URLs and saving those [14].

- **Ordering Time:** Time required to Order URLs based on Content and Structural Similarity.
- **Precision:** It is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.
- **Similarity Scores**: all the three similarity score- content, structural and total similarity scores.

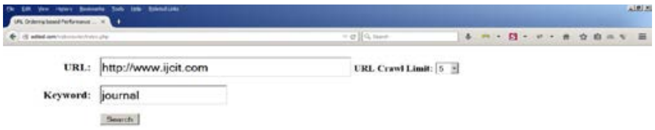
A. *Preparing Cases*

From Table I each international journal is taken up one by one on basis of two cases, where Case 1 is URL Crawl limit set to 5 and Case 2 is URL Crawl limit set to 10. The URLs of journals work as seed URLs with a keyword 'journal'. The input & output of crawling and ordering of journal web sites with the criterion of selected crawl limit are shown below.

Journal Website 1: http://www.ijcit.com - International Journal of Computer and Information Technology (IJCIT)

Case 1: URL Crawl Limit= 5

Figure 2 shows the input URL with keyword journal and URL Crawl Limit= 5. Figure 3 shows the output, where keyword journal is found in all 5 URLs. The content similarity score calculated on basis of similar content between pages with respect to keyword, is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.



URL: http://www.ijcit.com URL Crawl Limit: 5

Keyword: journal

Figure 2. Journal Website 1 with Crawl Limit 5

Case 2: URL Crawl Limit= 10

Figure 4 shows the input URL with keyword journal and URL Crawl Limit= 10. Figure 5 shows the output, where keyword journal is found in all 10 URLs. The content similarity score calculated on basis of similar content between pages with respect to keyword, is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.

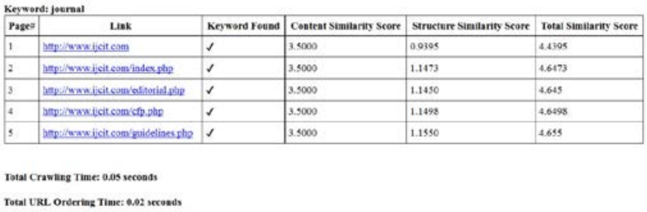
Journal Website 2: http://www.thesai.org - The Science and Information (SAI) Organization

Case 1: URL Crawl Limit= 5

Figure 6 shows the input URL with keyword journal and URL Crawl Limit= 5. Figure7 shows the output where keyword journal is found in all 4 URLs. The calculated content similarity score on basis of similar content between pages with respect to keyword is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.

Case 2: URL Crawl Limit= 10

Figure 8 shows the input URL with keyword journal and URL Crawl Limit= 10. Figure 9 shows the output where keyword journal is found in all 8 URLs. The content similarity score calculated on basis of similar content between pages with respect to keyword, is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.



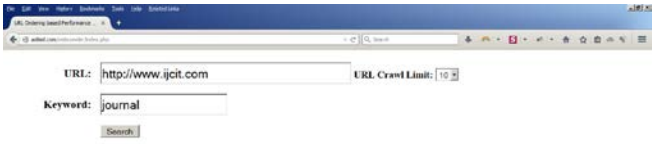
Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijcit.com	✓	3.5000	0.9395	4.4395
2	http://www.ijcit.com/index.php	✓	3.5000	1.1473	4.6473
3	http://www.ijcit.com/editorial.php	✓	3.5000	1.1450	4.645
4	http://www.ijcit.com/cfp.php	✓	3.5000	1.1498	4.6498
5	http://www.ijcit.com/guidelines.php	✓	3.5000	1.1550	4.655

Total Crawling Time: 0.05 seconds

Total URL Ordering Time: 0.02 seconds

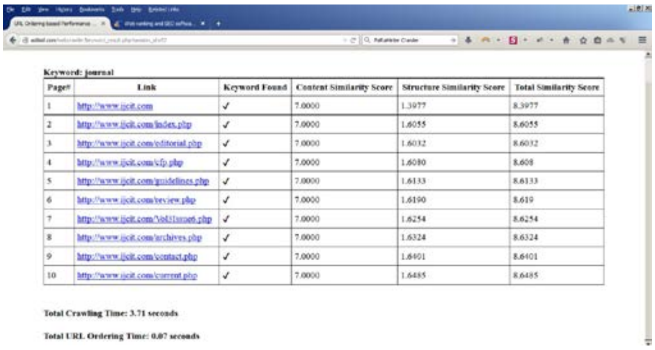
Figure 3. Results of Journal Website 1 with Crawl Limit 5



URL: http://www.ijcit.com URL Crawl Limit: 10

Keyword: journal

Figure 4. Journal Website 1 with Crawl Limit 10



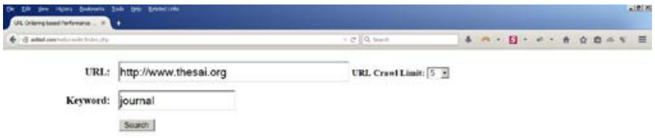
Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijcit.com	✓	7.0000	1.3977	8.3977
2	http://www.ijcit.com/index.php	✓	7.0000	1.6055	8.6055
3	http://www.ijcit.com/editorial.php	✓	7.0000	1.6032	8.6032
4	http://www.ijcit.com/cfp.php	✓	7.0000	1.6080	8.608
5	http://www.ijcit.com/guidelines.php	✓	7.0000	1.6133	8.6133
6	http://www.ijcit.com/review.php	✓	7.0000	1.6190	8.619
7	http://www.ijcit.com/Vol1Issue6.php	✓	7.0000	1.6254	8.6254
8	http://www.ijcit.com/archives.php	✓	7.0000	1.6324	8.6324
9	http://www.ijcit.com/contact.php	✓	7.0000	1.6401	8.6401
10	http://www.ijcit.com/current.php	✓	7.0000	1.6485	8.6485

Total Crawling Time: 3.71 seconds

Total URL Ordering Time: 0.07 seconds

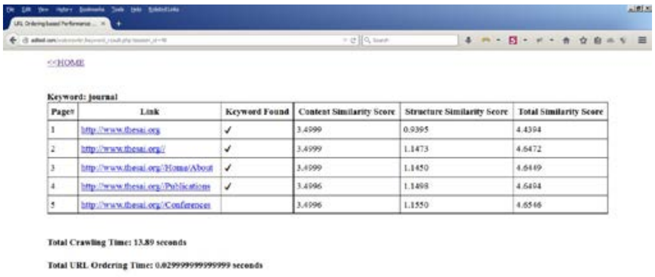
Figure 5. Results of Educational Website 1 with Crawl Limit 10



URL: http://www.thesai.org URL Crawl Limit: 5

Keyword: journal

Figure 6. Journal Website 2 with Crawl Limit 5



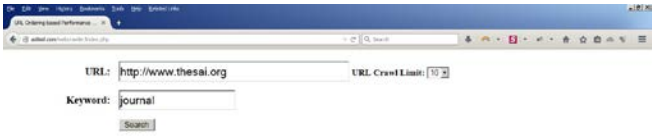
Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.thesai.org	✓	3.4999	0.9395	4.4394
2	http://www.thesai.org//	✓	3.4999	1.1473	4.6472
3	http://www.thesai.org//Home/About	✓	3.4999	1.1450	4.6449
4	http://www.thesai.org//Publications	✓	3.4996	1.1498	4.6494
5	http://www.thesai.org//Conferences		3.4996	1.1550	4.6546

Total Crawling Time: 13.89 seconds

Total URL Ordering Time: 0.029999999999999 seconds

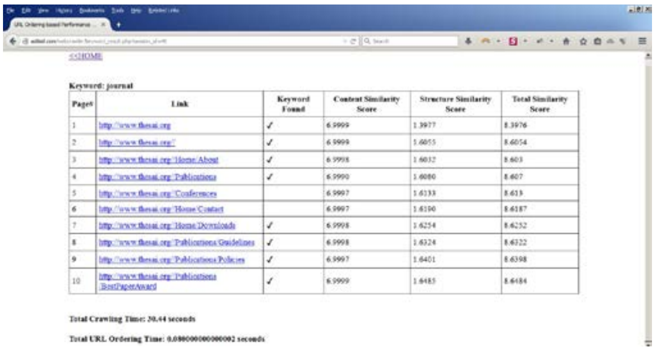
Figure 7. Results of Journal Website 2 with Crawl Limit 5



URL: http://www.thesai.org URL Crawl Limit: 10

Keyword: journal

Figure 8. Journal Website 2 with Crawl Limit 10



Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.thesai.org	✓	6.9999	1.3977	8.3976
2	http://www.thesai.org/	✓	6.9999	1.6055	8.6054
3	http://www.thesai.org/Home/About	✓	6.9998	1.6032	8.603
4	http://www.thesai.org/Publications	✓	6.9990	1.6080	8.607
5	http://www.thesai.org/Conferences		6.9997	1.6133	8.613
6	http://www.thesai.org/Home/Contact		6.9997	1.6190	8.6187
7	http://www.thesai.org/Home/Downloads	✓	6.9998	1.6254	8.6252
8	http://www.thesai.org/Publications/Guidelines	✓	6.9998	1.6324	8.6322
9	http://www.thesai.org/Publications/Policies	✓	6.9997	1.6401	8.6398
10	http://www.thesai.org/Publications/BestPaperAward	✓	6.9999	1.6485	8.6484

Total Crawling Time: 20.44 seconds

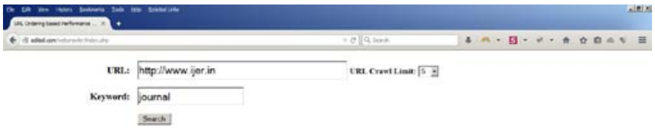
Total URL Ordering Time: 0.080000000000002 seconds

Figure 9. Results of Journal Website 2 with Crawl Limit 10

Journal Website 3: http://www.ijer.in - International Journal of Engineering Research (IJER)

Case 1: URL Crawl Limit= 5

Figure 10 shows the input URL with keyword journal and URL Crawl Limit= 5. Figure 11 shows the output where keyword journal is found in only 2 URLs. The content similarity score calculated on basis of similar content between pages with respect to keyword, is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.



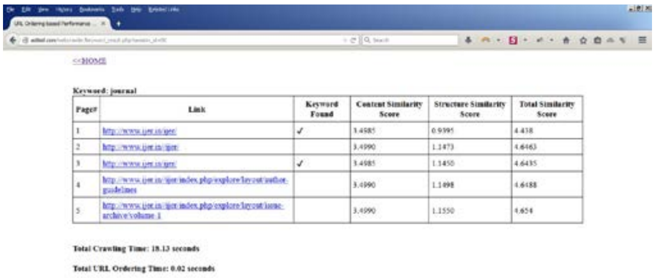
URL: http://www.ijer.in URL Crawl Limit: 5

Keyword: journal

Figure 10. Journal Website 3 with Crawl Limit 5

Case 2: URL Crawl Limit= 10

Figure 12 shows the input URL with keyword journal and URL Crawl Limit= 10. In Figure 13, keyword journal is found in only 2 URL. The content similarity scores calculated on basis of similar content between pages with respect to keyword, and Structural similarity scores based on the linking structure of these hyperlinks with total similarity scores are shown.



Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijer.in/ijer/	✓	3.4985	0.9395	4.438
2	http://www.ijer.in/ijer/		3.4990	1.1473	4.6463
3	http://www.ijer.in/ijer/	✓	3.4985	1.1450	4.6435
4	http://www.ijer.in/ijer/index.php/explore/layout/author-guidelines		3.4990	1.1498	4.6488
5	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-1		3.4990	1.1550	4.654

Total Crawling Time: 18.13 seconds

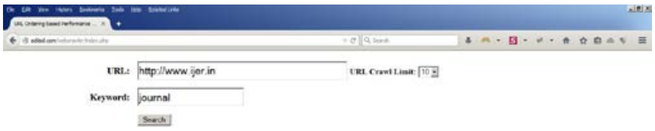
Total URL Ordering Time: 0.02 seconds

Figure 11. Results of Journal Website 3 with Crawl Limit 5

Journal Website 4: http://www.ijsce.org - International Journal of Soft Computing and Engineering (IJCSE)

Case 1: URL Crawl Limit = 5

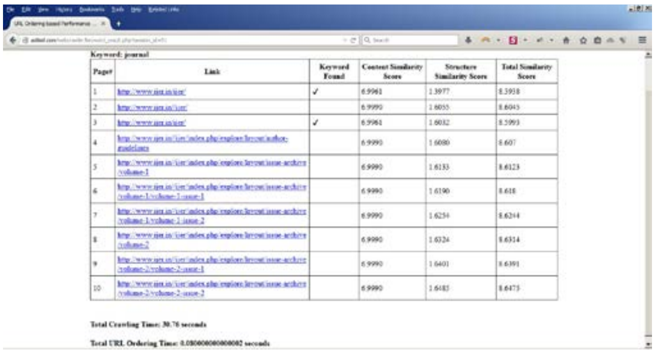
Figure 14 shows the input URL with keyword journal and URL Crawl Limit= 5. Figure 15 shows the output where keyword journal is found in all 5 URLs. The calculated content similarity score on basis of similar content between pages with respect to keyword is shown. Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.



URL: http://www.ijer.in URL Crawl Limit: 10

Keyword: journal

Figure 12. Journal Website 3 with Crawl Limit 10



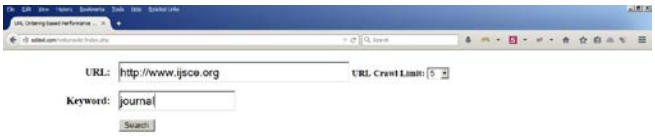
Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijer.in/ijer/	✓	6.9961	1.3977	8.3938
2	http://www.ijer.in/ijer/		6.9990	1.6055	8.6045
3	http://www.ijer.in/ijer/	✓	6.9961	1.6032	8.5993
4	http://www.ijer.in/ijer/index.php/explore/layout/author-guidelines		6.9990	1.6080	8.607
5	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-1		6.9990	1.6133	8.6123
6	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-1/volume-1-issue-1		6.9990	1.6190	8.618
7	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-1/volume-1-issue-2		6.9990	1.6254	8.6244
8	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-2		6.9990	1.6324	8.6314
9	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-2/volume-2-issue-1		6.9990	1.6401	8.6391
10	http://www.ijer.in/ijer/index.php/explore/layout/issue-archive/volume-2/volume-2-issue-2		6.9990	1.6485	8.6475

Total Crawling Time: 30.78 seconds

Total URL Ordering Time: 0.080000000000002 seconds

Figure 13. Results of Journal Website 3 with Crawl Limit 10



URL: http://www.ijsce.org URL Crawl Limit: 5

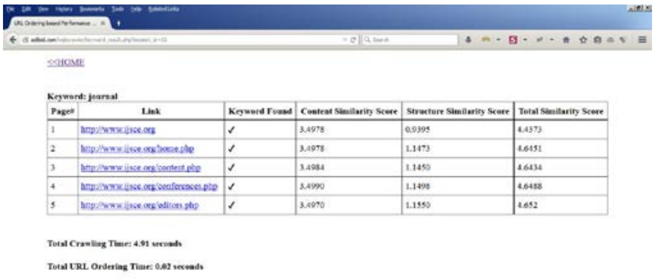
Keyword: journal

Figure 14. Journal Website 4 with Crawl Limit 5

Case 2: URL Crawl Limit= 10

Figure 16 shows the input URL with keyword journal and URL Crawl Limit= 10. Figure 17 shows the output where keyword journal is found in all 10 URLs. The calculated content similarity score on basis of similar content between pages with respect to keyword is shown.

Structural similarity based on the linking structure of these hyperlinks with total similarity score is also shown.



<<HOME

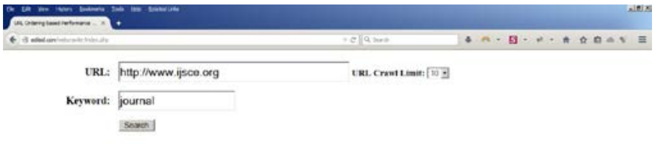
Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijsce.org	✓	3.4978	0.9395	4.4373
2	http://www.ijsce.org/home.php	✓	3.4978	1.1473	4.6451
3	http://www.ijsce.org/content.php	✓	3.4984	1.1450	4.6434
4	http://www.ijsce.org/conferences.php	✓	3.4990	1.1498	4.6488
5	http://www.ijsce.org/editors.php	✓	3.4970	1.1550	4.652

Total Crawling Time: 4.91 seconds

Total URL Ordering Time: 0.02 seconds

Figure 15. Results of Journal Website 4 with Crawl Limit 5

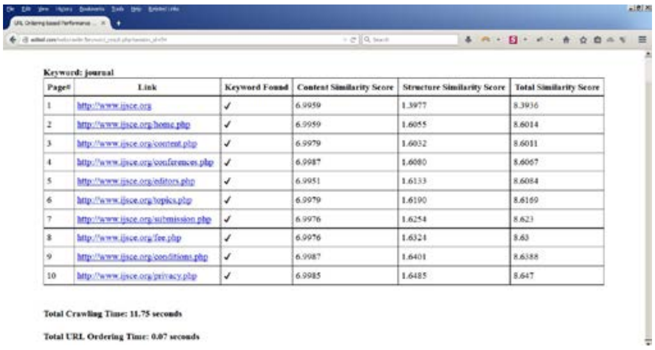


URL: http://www.ijsce.org

URL Crawl Limit: 10

Keyword: journal

Figure 16. Journal Website 4 with Crawl Limit 10



Keyword: journal

Page#	Link	Keyword Found	Content Similarity Score	Structure Similarity Score	Total Similarity Score
1	http://www.ijsce.org	✓	6.9959	1.3977	8.3936
2	http://www.ijsce.org/home.php	✓	6.9959	1.6055	8.6014
3	http://www.ijsce.org/content.php	✓	6.9979	1.6032	8.6011
4	http://www.ijsce.org/conferences.php	✓	6.9987	1.6080	8.6067
5	http://www.ijsce.org/editors.php	✓	6.9951	1.6133	8.6084
6	http://www.ijsce.org/topics.php	✓	6.9979	1.6190	8.6169
7	http://www.ijsce.org/submission.php	✓	6.9976	1.6254	8.623
8	http://www.ijsce.org/fee.php	✓	6.9976	1.6324	8.63
9	http://www.ijsce.org/conditions.php	✓	6.9987	1.6401	8.6388
10	http://www.ijsce.org/privacy.php	✓	6.9985	1.6485	8.647

Total Crawling Time: 11.75 seconds

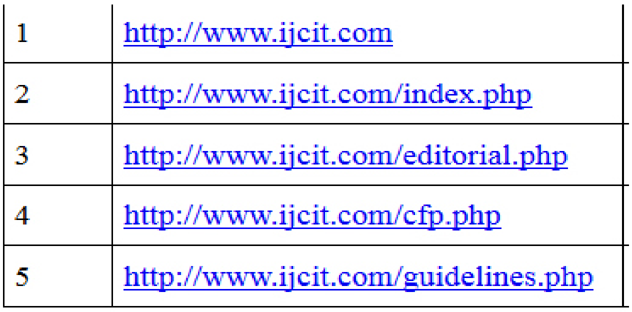
Total URL Ordering Time: 0.07 seconds

Figure 17. Results of Journal Website 4 with Crawl Limit 10

B. *Running Parameters*

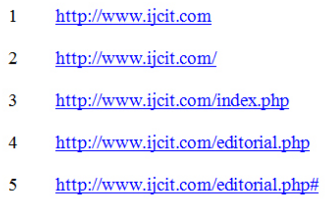
1) Top URLs: Comparison is done on top five URLs crawled by implemented web crawler and PageRank based - Parameter v1.4.8 (developed by Cleverstat.com).

Figure 18 shows the crawled URL given as output by developed crawler while Figure 19 shows URLs crawled by PageRank based - Parameter v1.4.8. Figure 18 clearly depicts that developed web crawler provides better and more unique URLs than PageRank based – Parameter shown in Figure 19.



1	http://www.ijcit.com
2	http://www.ijcit.com/index.php
3	http://www.ijcit.com/editorial.php
4	http://www.ijcit.com/cfp.php
5	http://www.ijcit.com/guidelines.php

Figure 18. Top URLs given by Developed Crawler



1	http://www.ijcit.com
2	http://www.ijcit.com/
3	http://www.ijcit.com/index.php
4	http://www.ijcit.com/editorial.php
5	http://www.ijcit.com/editorial.php#

Figure 19. Top URLs given by PaRaMeter Crawler

2) Crawling Time: Time to crawl each of the all four sites is compared when URL Limit is 5 and10 and keyword is journal. Table II provides the crawling time of all four websites when crawl limit is 5 and 10 URLs and Figure 20 depicts the crawling time graphically.

It can be concludes that when crawl limit is 5 then the URLs in increasing order of their crawling time are: http://www.ijcit.com > http://www.ijsce.org > http://www.thesai.org > http://www.ijer.in

TABLE II. CRAWLING TIME OF WEBSITES

	URL	Total Crawling Time (ms)	
		Crawl Limit 5	Crawl Limit 10
1	http://www.ijcit.com	50	3710
2	http://www.thesai.org	13890	30440
3	http://www.ijer.in	18130	37760
4	http://www.ijsce.org	4910	11750

When crawl limit is 10 then their order is : http://www.ijcit.com > http://www.ijsce.org > http://www.thesai.org > http://www.ijer.in. It can be concluded that IJER website http://www.ijer.in take maximum crawling time.

TABLE III. ORDERING TIME OF WEBSITES

	URL	Total Ordering Time (ms)	
		Crawl Limit 5	Crawl Limit 10
1	http://www.ijcit.com	20	70
2	http://www.thesai.org	20	80
3	http://www.ijer.in	20	80
4	http://www.ijsce.org	20	70

3) Ordering Time: Time to order all four sites when keyword is journal and URL Limit is 5 and 10. Table 4.3 shows ordering time taken to order all four websites when URL Limit is 5 and 10. Figure 21 depicts ordering time graphically.

It can be concluded that the ordering time of websites in increasing order when URL crawl limit = 5 is http://www.ijcit.com = http://www.ijsce.org = http://www.thesai.org = http://www.ijer.in.

While the ordering time when crawl limit = 10 is http://www.ijcit.com > http://www.ijsce.org > http://www.thesai.org > http://www.ijer.in.

4) Precision: As per precision discussed above, it is calculated for the number of relevant and irrelevant documents retrieved multiplied by 100 to give a percentage when keyword is: journal.

The metric Precision which calculates percentage of relevant page, points out that when keyword journal is searched on the website http://www.ijer.in, it has least value, while http://www.ijcit.com and http://www.ijsce.org have the most relevant pages to the keyword journal. Table IV shows precision percentages for each website.

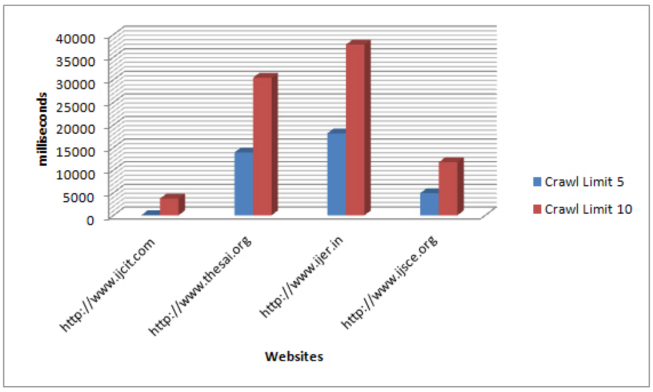


Figure 20. Graph depicting Crawling Time of web sites

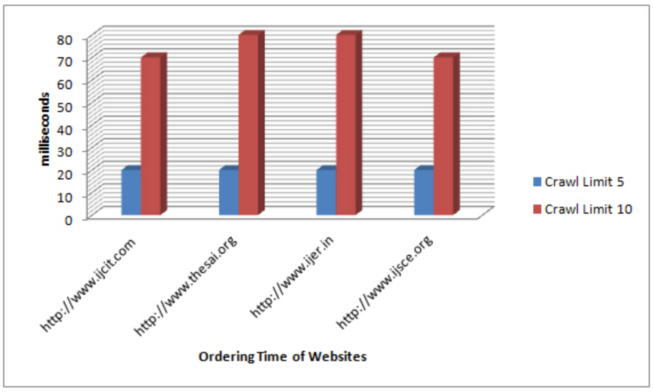


Figure 21. Graph depicting Ordering Time of web sites

5) Similarity Scores: Table V shows the similarity score of each web site when URL limit is 5 and 10 respectively. Figure 22, 23 and 24 depicts graphically content similarity, structural similarity and total similarity scores respectively.

As we can see in Table V, the content similarity score of URLs http://www.ijcit.com & http://www.ijsce.org are approximately same therefore their plotting in graph overlap. Thus both URLs http://www.ijcit.com & http://www.ijsce.org are depicted by only one plotted line in Figure 22.

TABLE IV. PRECISION PERCENTAGE WHEN KEYWORD IS JOURNAL

	URL	Cases	Precision (%)
1	http://www.ijcit.com	Crawl Limit 5	100
		Crawl Limit 10	100
2	http://www.thesai.org	Crawl Limit 5	80
		Crawl Limit 10	80
3	http://www.ijer.in	Crawl Limit 5	40
		Crawl Limit 10	20
4	http://www.ijsce.org	Crawl Limit 5	100
		Crawl Limit 10	100

TABLE V. SIMILARITY SCORES

	URL	Cases	Content Similarity Score	Structural Similarity Score	Total Similarity Score
1	http://www.ijcit.com	Crawl Limit 5	3.4998	0.2438	3.7436
		Crawl Limit 10	6.9994	0.2678	7.2672
2	http://www.thesai.org	Crawl Limit 5	3.5	0.2438	3.7438
		Crawl Limit 10	6.3	0.2630	6.563
3	http://www.ijer.in	Crawl Limit 5	1.25	0.2438	0.2438
		Crawl Limit 10	6.9992	0.2678	7.2670
4	http://www.ijsce.org	Crawl Limit 5	3.4999	0.2438	3.7437
		Crawl Limit 10	6.9999	0.2678	7.2677

As in Table V the structural similarity score of URLs http://www.ijcit.com, http://www.ijsce.org and http://www.ijer.in are approximately same therefore their plotting in graph overlap. Thus, URLs http://www.ijcit.com, http://www.ijsce.org and http://www.ijer.in are depicted using only one plotted line in Figure 23.

As it can be seen in Table V the total similarity score of structural similarity score of URLs http://www.ijcit.com and http://www.ijsce.org are approximately same due to similarity in content similarity score and structural similarity score, therefore their plotting in graph overlap. Thus both URLs http://www.ijcit.com and http://www.ijsce.org are depicted by only one plotted line in Figure 24.

IV. Conclusion and Future Work

In this work a web crawler was developed which used content and structural similarity of web pages to order them. The content similarity is calculated on basis of frequency of keyword among the crawled pages, while the structural similarity is calculated by pairing in-neighbors to a node, as defined in SimRank [8] algorithm. The crawler eliminates duplicate URLs, thus providing with unique URLs. A set of similar websites were given as input to crawler to crawl and the outputs were compared on the set of parameters such as top URLs, precision, crawling time, ordering time and similarity scores. The developed Web crawler shows web pages that are relevant to a query on basis of content and structural similarity.

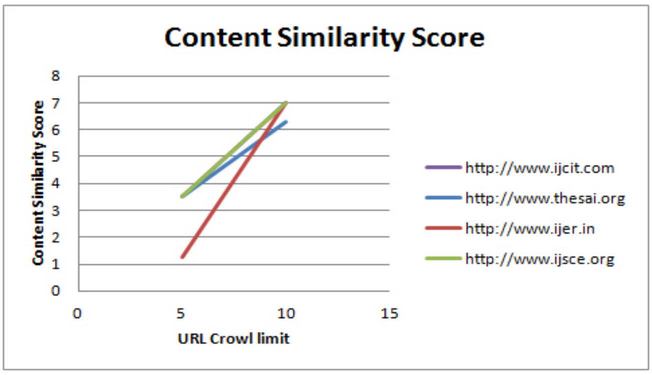


Figure 22. Graph for Content Similarity Score

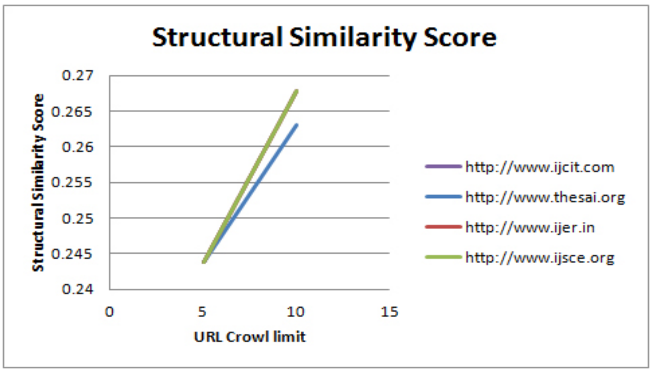


Figure 23. Graph for Structural Similarity Score

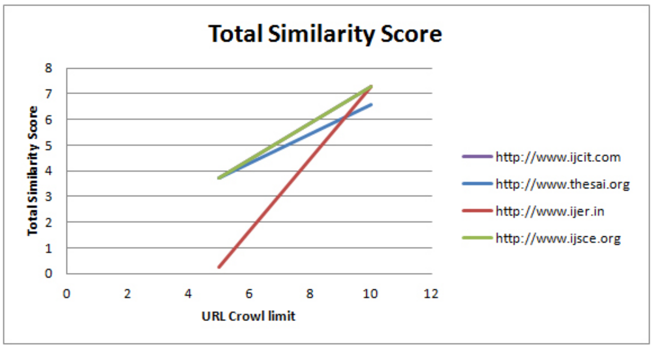


Figure 24. Graph for Total Similarity Score

The results of crawler can be made more relevant by using usage mining to calculate page popularity. A policy to check the frequency of revisiting URLs and then crawling fresh pages again can be implemented. Algorithm to judge changes in content and text in a web page can also be implemented. Crawler can be made more polite while crawling by using delay.

Acknowledgment

I express my deep gratitude to Asst. Prof. M. Akbar, Integral University, India for his helpful insight and generosity in terms of interactions, guidance and advices.

References

[1] KethiReddy, Thapar University, PhD. Thesis, Improving Efficiency of Web Crawler Algorithm Using Parametric Variations, 2010. Available at: http://dspace.thapar.edu:8080/dspace/handle/10266/1364.

[2] B. Liu, Web Data Mining- Exploring Hyperlinks, Contents and Usage Data, Springer-Verlag Berlin Heidelberg, 2007.

[3] D. Jiang, J. Pei, and H. Li, “Mining search and browse logs for web search: a survey”, ACM Transactions on Computational Logic, Vol. V, No. N, pages 1–42, 2013.

[4] L. Getoor, “Link mining: a new data mining challenge”, ACM SIGKDD Explorations, volume 5, issue 1, pages 84-93, July 2003.

[5] S. Brin, and L. Page, “The anatomy of a large-scale hypertextual Web search engine”, Computer Networks and ISDN Systems, Vol. 30, No.1-7, pp. 107-117, 1998.

[6] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment”, In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 668-677, 1998.

[7] C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti, “Social semantic query expansion”, ACM Transactions on Intelligent Systems and Technology (TIST) - Survey papers, special sections on the semantic adaptive social web, intelligent systems for health informatics, regular papers, vol. 4, issue 4, Article No. 60, September 2013.

[8] K.S. Kim, K.Y. Kim, K.H. Lee, T.K. Kim, and W.S. Cho, “Design and implementation of web crawler based on dynamic web collection cycle”, International Conference on Information Networking (ICOIN), IEEE, pp. 562 – 566, 2012.

[9] Yang Sun, I.G. Councill, and C.L. Giles, “The ethicality of Web crawlers”, International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, vol. 1, pp.668 – 675, 2010.

[10] S. Qiao, T. Li, and J. Qiu, “SimRank: A Page Rank approach based on similarity measure”, Intelligent System and Knowledge Engineering, IEEE, pp 390-395, 2010.

[11] G. Salton, and C. Buckley, “Term-weighting approaches in automatic text retrieval”, Information Processing & Management, vol. 24, issue 5, pp. 513-523, 1988.

[12] G. Jeh, and J. Widom, “SimRank: a measure of structural-context similarity”, Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 538-543, 2002.

[13] J. Cho, H. Garcia-Molina, and L. Page, “Efficient crawling through URL ordering”, Journal Computer Networks and ISDN Systems, Vol. 30, Issue 1-7, pp. 161-172, Elsevier Science Publishers, 1998.

[14] Q. Tan and P. Mitra, “Clustering-based incremental web crawling”, ACM Transaction on Information Systems, vol. 28, issue 4, Article 17 November 2010.

[15] L, K. Soon, K. B. Hwang, and S. H. Lee, “An empirical study on harmonizing classification precision using IE patterns”, 2nd International Conference on Software Engineering and Data Mining (SEDM), IEEE, pp. 251-256, 2010.