A NOVEL TECHNIQUE FOR BACK-LINK EXTRACTION AND RELEVANCE EVALUATION

Pooja Gupta[1] and A. K. Sharma[2] and Divakar Yadav[3]

[1]Department of Computer Science Engineering, MAIT, Rohini, New Delhi
poojaguptamait@gmail.com
[2]Department of Computer, YMCA University of Science and Technology, Faridabad
ashokkale2@rediffmail.com
[3]Department of Computer Science and Information, JIITU, Noida
divakar.yadav@jiit.ac.in

ABSTRACT

Focused Crawling aims to search the WWW in relevance to the user topic of interest, tends to provide more number of relevant documents in first top results, leading to the need of maintaining the index with more number of related web pages; using the relative measure of relevancy between the documents. This paper provides a novel technique for extracting the back-links of a web page and to evaluate their context score; that helps to update the document index with more number of relevant documents.

KEYWORDS

Context Score, Back-link, Relevant Pages, Common Keywords, WWW

1. INTRODUCTION

With the exponential growth of WWW and the huge amount of information available on it, the size of indexes maintained by all existing search engines has also become massive consisting of enormous entries. Therefore the results provided by a Search Engine in response to a query may contain several thousands or millions of references to web pages. Many of the web pages that are returned are of low quality much against the desired topic of interest. Whereas, user is only interested to see first few relevant results. So, this is a challenge for search engine how to index the higher quality web pages and also to place them on the top most position of the result list.

Web structure plays potential role in the evaluation of web pages. Hyperlinks present in a web page provide mean to measure it. The various types of links a web page contains are internal, external, transverse and intrinsic links. Generally the relevancy of documents returned by the search engine is computed on the basis their link analysis i.e. higher the number of hyperlinks to/from that page; more relevant that document will be considered. Thus it is very important to find out the forward link as well as the backward link for all the web documents in order to compute their relevance. Further, in order to find out the web documents relevant to the users query, it is assumed that if a particular web document is relevant to the user's query then its preceding page may also be important in order to satisfy user's query requirement. For example, if a user is searching for query "M.Tech" and clicking on a particular document from the results returned by the search engine, then its preceding page that contains "Courses" will also be important for the user. In this paper a novel technique to find out the back-links of the web pages is being introduced so that relevance of the web pages pointed by those back-links can be evaluated.

The paper has been organized as follows Section 2 illustrate Related Work, Section 3 is explaining the Proposed Back Link Extraction and Relevance Evaluation Technique; where Section 3.1 describe the algorithm steps to compute context score of a back-link. Section 4

DOI : 10.5121/ijcsit.2011.3316

shows the implementation/experimental results with the screen shots and respective data entries in databases. Section 5 presents the conclusion and section 6 layout the references.

2. RELATED WORK

Generally there is very less information available related to the extraction of back-pointers of a web page. This effect the web surfing and information sighting. Related web documents that are hyperlinked to a web page are not located at a single place. S. Chakarbarti [17] found that if back-link information will be provided to web surfer; the process of information sighting will be much more effective.

The hyperlink structure of the WWW is one of its important significant characteristics. The hyperlink structure plays potential role in evaluating a web page relevancy; that helps the search engine to take decision for satisfying the user query. Existing studies on various focused/topical crawlers by various researchers use the entire content of the Web page to evaluate the context of the hyperlink in that Web page [1, 2, 3 and 4]. Some researchers have discussed techniques that select a few words around the hyperlink, as the link context [5, 6 and 7].

Page rank score [9, 10, 11 and 12] is calculated based on the number of back-links a web page has and the popularity of that page. The page rank algorithm first assigns a manual score to the initial small set of web pages then starts following the hyperlinks between these web pages and calculate the score for each new page depending on the number of back-link that new page has. More the number of back-links more popular that page is considered. The page rank algorithm is processed on all documents. HITS algorithm [19] assigns authority and hub score to each page depending on the query keywords. Where, hub score of a page is with number of links to other pages and authority score of a page is the number of links points to that page by different hubs. Both the page rank and HITS algorithm consider all hyperlinks equally important irrespective of their context. HITS algorithm works at the query evaluation time not at the time of index creation. Thus, authority and hub score are query dependent.

Guang in [16] proposed a level-based link analysis that computes the rank of a web page by assigning weight to each hyperlink according to its level properties. A link analysis page rank algorithm that works on back-link count and association metric to evaluate relevancy of a link in a web page before the actual crawling is recommended by S. Ganesh [15]. Another researcher has used the hyperlink anchor text to evaluate the context of the associated page in [18]. He has applied a filtering mechanism based on linguistic analysis of all context sentences to get the best illustrated context of associated page. Chen Ding [14] has proposed a mechanism to locate referral parent for a given fragment of a web document from the client side. The mechanism explores the hierarchical structure of a web document fragment.

The review of the available research indicates that the search engines suffer from the following drawbacks:

1. None of the search engine differentiates between various incoming links. Hence, all back-links are considered of equal importance.

2. While calculating the authority and hub pages search engine does not see to mutually reinforcing relationship. That is if two pages having lot of links pointing each other are not checked, that increase the authority and hub scores to those web pages.

3. Some web pages contain links pointing to irrelevant web pages.

4. Consider only number of in-links and out-links of a web page to score that page, but do not consider the contextual relevancy of pages linked to these links.

In the proposed back-link technique the relevancy of back-links are evaluated by extracting the keywords that back-link page contains and if that matches to the initial web page to some extend then place that URL in the repository under that topic of interest corresponding to the

initial web page to enrich the database in that specific topic that helps to answer similar query more efficiently and quickly in future.

3. PROPOSED BACK LINK EXTRACTION AND RELEVANCE EVALUATION TECHNIQUE

In the proposed technique it is being felt that the hyperlinks in a page play potential role to find out the more pages related to same topic. It is being noted that if a particular web-page is more relevant when evaluated against a user query the pages pointing to and from this page may also be relevant to the same topic or area. Davison [11] tested hypotheses related to topical locality of the web. Most web pages have links to other pages with similar context. So, a source page (here represented as Parent page) is the page where hyperlinks appears and hyperlinks pages are (taken as the child page) pages that the hyperlink leads to. The source page will become the back-link of the hyperlinks page.

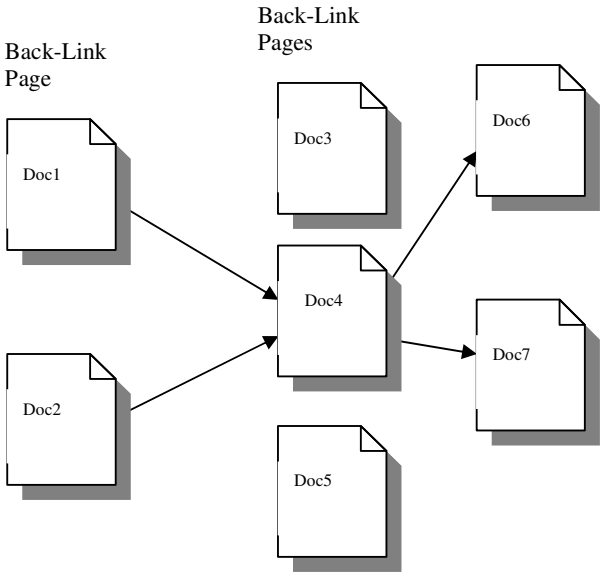

Figure 1. Back-Links

Consider Figure1. wherein Doc1, Doc2, Doc4 are back-links for Doc6 and Doc7 both. In this work a technique called 'Context Oriented Back-Links' based on back-link relevance evaluation is being proposed. The block diagram of architecture framework is as shown in the Figure 2.

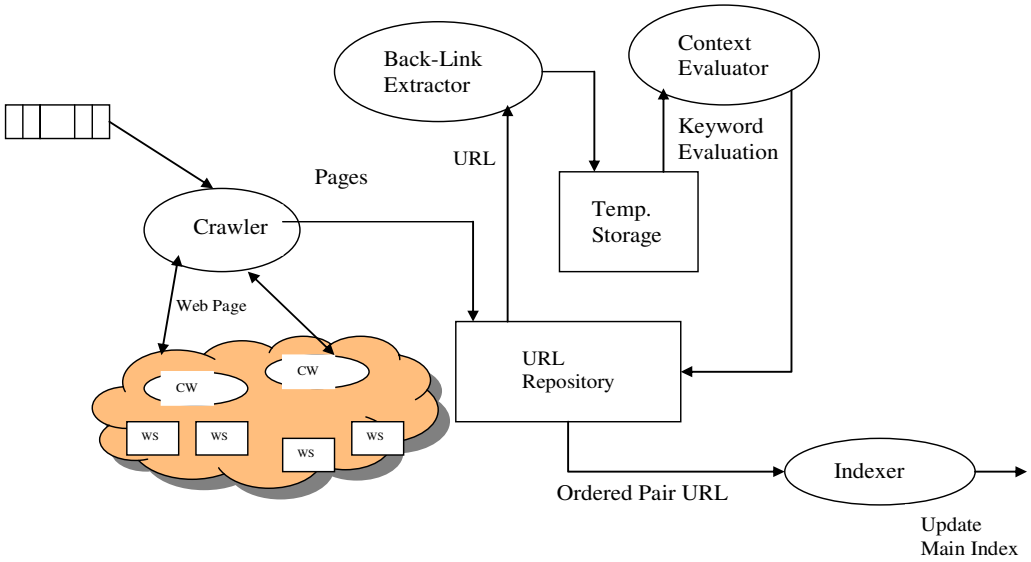

Figure 2. Architecture Framework for Back-link Extractor

The architecture framework contains three components Crawler, Back-Link Extractor and the Context Evaluator. Crawler component is the basic components that receive the seed URL and download the corresponding web page. It works recursively on downloaded pages to find the hyperlinks and stores the retrieved pages in the database i.e. URL Repository. Back-Link Extractor picks the URL pair from the URL Repository and then finds out the back-links of each URL; stores them in a temporary database. Context Evaluator is the component that takes the ordered pair back-link URLs from the temporary database, extracts keywords of the pages corresponding to back-link URLs using the web structure of the pages. It evaluates the similarity between back-link URLs Pages in terms of 'Context Score'. Once done for all the ordered pair URLs; eliminates the less contextual related pair and update the URL repository with more contextual related pair information; that passed to the 'Indexer' to endeavour this information while updating the main Index.

Thus, all the links to and from a URL are not considered of equal importance as done by other existing search engines, moreover evaluation is done on the basis of context. Hence, the user is served with more context oriented results for a given query. The brief functionality of the architecture components is illustrated in Table 1.

Table 1. Component and their functionality

Component	**Functionality**
Crawl Manager	Core component works on the list of seed URLs. It distributes these URLs to multiple crawl workers for downloading. It receives the downloaded web documents from them and stores them in the local database.
Crawl Worker	This component is under the control of crawl manager. It downloads the web documents for the list of URLs assigned by Crawl manager and repeats the process recursively.
Back-link Extractor	The Back-link extractor takes the URLs from the database and processes them to give all possible ordered pair of URLs present in the same hierarchy.
Context Evaluator	The context evaluator evaluates the context score by finding the number of common meta keywords between the parents and the child URLs.

The main data structure of the system is URL table (URL_Information). The structure of the URL table is as shown in Table 2.

Table 2. Structure of the URL table

URL	**Hyperlink**
URL 1	URL 2
URL 2	URL 3
URL 3	URL 4
X	Y
X	Z
X	V
Y	B
B	Z
P	Q
…….	……..

The structure comprises of two fields 'URL' and 'Hyperlink'. Where URL is the URL of any web page and Hyperlink is the corresponding hyperlink from that URL , if multiple hyperlinks are present in a specific URL, multiple entries are done for each hyperlink as in the Table 2.multiple entries done for URL 'X' and its corresponding hyperlinks.

3.1. Algorithm Context Score (URLs)

// This algorithm extract the back-links and evaluates the Context Score//

Begin

Step 1. For each URL 'i' in the database, URL table is searched to find a match with Hyperlink say 'j'.

2. If a match is found the corresponding (i, j) row is selected, where 'j' will be the back-link of 'i'

3. From the ordered pair (i, j) now the value 'j' is searched recursively in Hyperlink field till it results in a match

4. All the ordered pair (i, j) entries of URL and corresponding Hyperlinks thus obtained by step 3 are stored and then passed to Crawler to download the corresponding web pages.

5. For each ordered pair entry (i, j) corresponding downloaded web pages are processed to find the total no. of keywords and the no. of common keywords.

6. The context score of the downloaded web page 'j' in respect to web page 'i' is calculated as –

   Context Score (CS) = $K[i] \cap K[j] / \sum (K[i])$

   K [i] is the set of keywords present in web page corresponding to URL 'i'

   K [j] is the set of keywords present in web page corresponding to URL 'j'

End;

The web page corresponding to URL 'j' is considered to be of high significance w.r.t web page corresponding to URL 'i' if it has a Context Score value higher than the value say 'α'. Here, 'α' is considered equal to 0.2(20% similarity). URL 'j' is the back-link of URL 'i'.

Thus, with the help of Back-link extractor and Context Evaluator the Indexer enrich the Main Index with more number of contextually related documents.

Hence, when a query is solved using this index, search engine results with more related web pages in top list to satisfy the user need of information in that specific topic.

4. EXPERIMENTAL RESULTS

The implementation of above architecture is done in Java connected with Oracle 10g Express. Several experiments have been done to find out the performance of the proposed system.

The Crawler starts with a seed URL list consisting of 15 initial URLs and it crawled about 2878 pages. It works recursively on the downloaded pages to find in-links and stores the retrieved URLs in the table named URL table. The back-link extractor takes the URLs from this table and processes them to give all ordered pair of URLs present in the same hierarchy and stores them in table named BTR. It has been found that the numbers of entries in BTR table corresponding to all crawled pages are about 1423. The Context evaluator evaluates the Context Score and update the table BTR. Finally, the FINAL_TABLE is created that contains only those pair of URLs having Context Score more than 'α'. It has been observed that corresponding to 1423 ordered pair URLs in BTR only 1283 are with significant Context Score.

The algorithm has been implemented and the user interface designed for the same is as shown in the Figure 3.

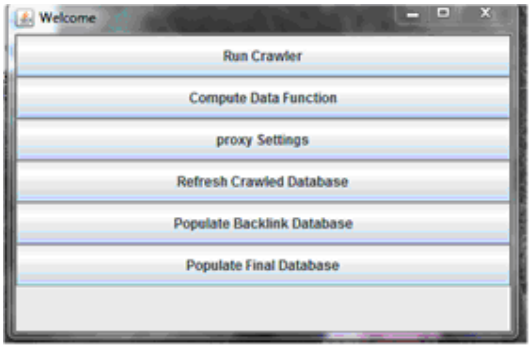

Figure 3. User Interface

The interface module comprises of six features as shown in table 3.

Table 3. User Interface Features

Feature	**Functionality**
Run Crawler	Web crawler that starts working starting with seed URL
Compute Data Function	compares the two Web pages corresponding to the two URLs mentioned and calculate the Context Score
Proxy setting	set the Proxy Server for the ON-Line connection required to run the crawler and other functions
Refreshed Crawled Database	Refresh the complete database
Populate Back Link Database	Computes the context score and populate the BTR table
Populate Final Database	Analysis the BTR table for records having CS >=0.2 and update the Final Table

4.1. Web Crawler

It starts working with list of seed URLs, extracts the hyperlinks recursively from each URL and stores them in a table. The figure 4 shows the screen shot of Web Crawler for the seed URL: "http://www.yahoo.com"

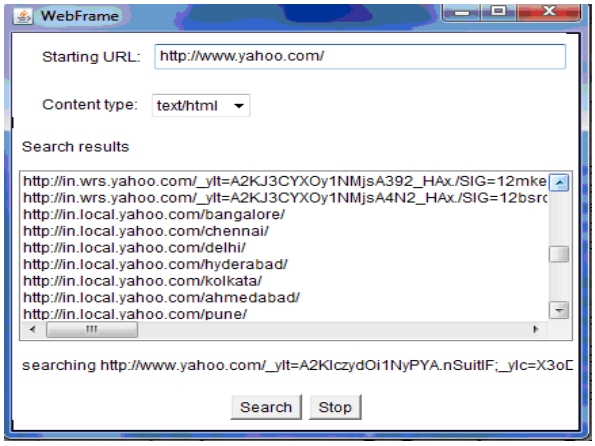

Figure 4. Web Crawler

4.2. Compute Data Function

It compares the two given URLs given in options for Website 1 and Website 2 and results in number of common keywords in both the URLs, total no of keywords in Website1 and the calculated Context Score. The list of different keywords in both the URLs is also shown in command line window at right side. Figure 5 shows the results for 2 given URLs

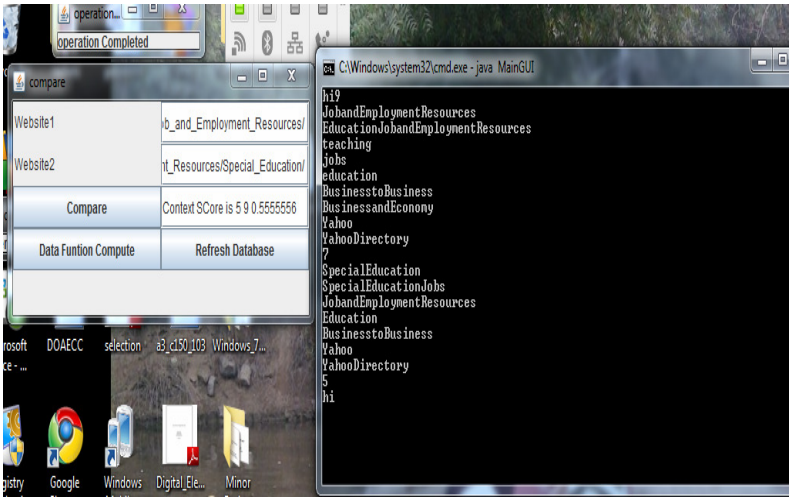

Figure 5. Compute Data Function

4.3. Proxy Settings

This feature is embedded to set IP address and proxy server to avail internet connection; required for the execution of whole module. Figure 6 shows the screen shot for these settings.

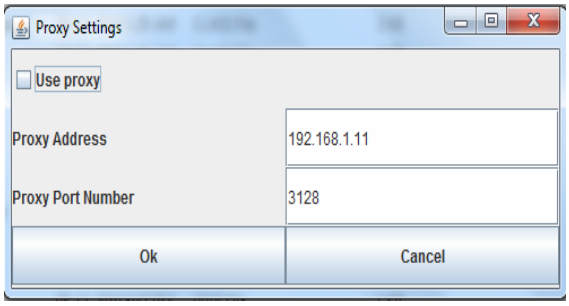

Figure 6. Proxy Setting

4.4. Back-link Tracking Records (BTR)

BTR table contains all ordered pair URLs extracted by back-link extractor with the number of common keywords, total number of keywords in child URL and the calculated Context Score. Table 4 shows some of the values.

Table 4. BTR values

S.No.	CHILD_URL	PARENT_URL	Total No. of Keywords	No. of Common Keywords	CONTEXTSCORE
1	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/English_as_a_Second_Language	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	9	5	0.55
2	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/K_12_Listings/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	9	5	0.55
3	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/Resumes/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	7	5	0.7

4	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/Special_Education/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	7	5	0.7
5	http://dir.yahoo.com/Education/Organizations/Professional/Unions/	http://dir.yahoo.com/Education/Organizations/Professional/	6	4	0.67
6	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/University_Listings/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	7	5	0.71
7	http://dir.yahoo.com/Education/Journals/	http://dir.yahoo.com/Education/	5	3	0.6
8	http://dir.yahoo.com/Education/Theory_and_Methods/Journals/	http://dir.yahoo.com/Education/Journals/	6	4	0.66
9	http://dir.yahoo.com/Education/History/Journals/	http://dir.yahoo.com/Education/Journals/	6	4	0.66
10	http://dir.yahoo.com/Education/Instructional_Technology/Journals/	http://dir.yahoo.com/Education/Journals/	6	4	0.66
11	http://dir.yahoo.com/Entertainment/Music/Education/Journals	http://dir.yahoo.com/Education/Journals/	7	4	0.57
12	http://www.tcrecord.org/	http://dir.yahoo.com/Education/Journals/	32	2	0.06

4.5. Contextually Significant Pairs

The BTR records are analysed to get more significant ordered pair of URLs. The ordered pair of URLs having context score more than 0.2 are considered to be more significant and are stored in a Final Table with their respective Context Score. Table 5 shows some of the results.

Table 5. Final Values

S. No.	CHILD_URL	PARENT_URL	CONTEXTSCORE
1	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/English_as_a_Second_Language/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	0.55
2	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/K_12_Listings/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	0.55
3	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/Resumes/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/Resumes/	0.7
4	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/Special_Education/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	0.7
5	http://dir.yahoo.com/Education/Organizations/Professional/Unions/	http://dir.yahoo.com/Education/Organizations/Professional/	0.67
6	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/University_Listings/	http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Education/Job_and_Employment_Resources/	0.71
7	http://dir.yahoo.com/Education/Journals/	http://dir.yahoo.com/Education/	0.6
8	http://dir.yahoo.com/Education/Theory_and_Methods/Journals/	http://dir.yahoo.com/Education/Journals/	0.66
9	http://dir.yahoo.com/Education/History/Journals/	http://dir.yahoo.com/Education/Journals/	0.66
10	http://dir.yahoo.com/Education/Instructional_Technology/Journals/	http://dir.yahoo.com/Education/Journals/	0.66
11	http://dir.yahoo.com/Entertainment/Music/Education/Journals/	http://dir.yahoo.com/Education/Journals/	0.57

5. CONCLUSIONS

Focused crawling aims to search only the relevant subset of the WWW for a specific topic of user interest. Whereas context focused crawler works to get contextually related documents to serve a user query in order to result with more number of relevant documents related to the user interest. It has been observed that if a document serves a user query well, its parent documents also serve well. So, back-links of URLs are considered important to get the more number of relevant documents for a given query. In addition, not all the back-link URLs are important. Thus, a technique to find out the back-links of a URL and then to find out the similarity between corresponding documents has been proposed. The Context Score is the measure that finds out the similarity between documents related to the URLs and its Back-links. The proposed technique first finds out the back-links of URLs and then eliminates the less contextually related pairs of URLs. Thus results in list of contextually more related ordered pair of URLs for future reference. It has been observed that final results contains only the ordered pair of URLs with significant high context score and less significant pairs of URLs having context score less than 0.2 has been eliminated. Thus only the important pairs of URLs are stored for the future reference to serve a query. Now, index is containing more relevant URLs related to the query at a single place in consecutive rows, thereof speeding up the search process.

In future the proposed back-link extraction module will be expanded to get the different senses of the keywords and then update the index with respect to different senses.

6. REFERENCES

[1] S. Chakarbarti, M. van den Berg & B. Dom, (1999) "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery" , Proc. Eighth Int'l World Wide Web Conf.

[2] P.M.E. De Bra & R.D.J. Post, (1994) "Information Retrieval in the World Wide Web: Making Client-Based Searching Feasible", Proc. First Int'l World Wide Web Conf.

[3] M. Diligenti, F. Coetzee, S. Lawrence, C.L. Giles & M. Gori, (2000) "Focused Crawling Using Context Graphs", Proc. 26th Int'l Conf. Very Large Data Bases , pp 527-534.

[4] G. Pant, K. Trioutsiouliklis, J. Jhonson & C. L. Giles, (2004) "Panorama: Extending Digital Libraries with Topical Crawlers", Proc. Fourth ACM/IEEE-CS Joint Conf. Digital libraries, pp 142-150.

[5] S. Chakarbarti, K. Punera & M. Subramanyam, (2002) "Accelerated Focused Crawling through Online Relevance Feedback", Proc. 11th Int'l World Wide Web Conf.

[6] M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalhaim & S. Ur (1998) "The Shark-Serach Algorithm –An Apllication: Tailored Web Site Mapping", Proc. Seventh Int'l World Wide Web Conf.

[7] G. Pant & F. Menczer (2003) " Topical Crawling for Business Intelligence", Proc. Seventh European Conf. Research and Advanced Technology for Digital Libraries.

[8] http://www.lib.berkely.edu/TeachingLib/Guides/Internet/SearchEngines.html (Reterievd: May 2010)

[9] Monika R. Henzinger, (2001) Hyperlink Analysis for the Web, IEEE Internet Computing, URL:http://maya.cs.depaul.edu/~classes/ds575/papers/hyperlink.pdf

[10] J. Han & K. C. C. Chang (2002) Data Mining for Web Intelligence, IEEE Computer Society, Vol. 35, Issue. 11, pp 64-70

[11] S. Brin & L. Page (1998) "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Seventh Int'l World Wide Web Conf.

[12] Phil Craven, Google's Page Rank Explained , Copyright WebWorkshop

[13] B. D. Davison (2000), "Topical Locality in the Web", Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval.

[14] Chen Ding, Chi-Hung Chi, Vincent Tam (2000) "Reverse Mapping of Referral Links from Storage Hierarchy for Web Documents", In Proceedings of ICTAI, pp 216~216

[15] S. Ganesh (2005) "Ontology Based Web Crawling – A Novel Approach", AWIC, LNAI 3528, pp 140-149.

[16] Guang Feng, Tie-Yan Liu, Xu-Dong Zhang, Tao Qin, Bin Gao, Wei-Ying Ma(2005),"Level-Based Link Analysis", In Proceedings of APWeb, pp.183~194.

[17] Chakrabarti S., Gibson, D. A., McCurley, K. S.(1999)," Surfing the Web Backwards", In the proceedings of 8th World Wide Web Conferences.

[18] N. Chauhan, A. K. Sharma (2008), " A framework to derive web pages context from hyperlink structure", IJICT 1(3/4), pp-329-346.

[19] Lecture No. 4, "HITS Algorithm – Hubs and Authorities on the Internet", http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

Authors

Prof. A. K. Sharma received his M.Tech. (Computer Sci. & Tech) with Hons. from University of Roorkee in the year 1989 and Ph.D (Fuzzy Expert Systems) from JMI, New Delhi in the year 2000. From July 1992 to April 2002, he served as Assistant Professor and became Professor in Computer Engg. at YMCA Institute of Engineering Faridabad in April 2002. He obtained his second Ph.D. in IT from IIIT & M, Gwalior in the year 2004. His research interests include Fuzzy Systems, Object Oriented Programming, Knowledge Representation and Internet Technologies

Pooja Gupta received the MCA degree with Gold Medal in 2002 and M.Tech degree with honours in Computer Science Engineering in 2006, both from Maharishi Dayanand University. Presently, she is working as a lecturer in Computer Science and Engineering Department in Maharaja Agrasen Institute of Technology (affiliated to I.P. University) Rohini, Delhi. She is also pursuing her Ph.D. in Computer Engineering and her areas of interests are Search Engines, Crawlers and Focused Crawling.

Dr. Divakar Yadav received his B.Tech. (Computer Science and Engineering) from Institute of Engineering & Technology, Lucknow, M.Tech. (Information Technology (Intelligent Systems)),in 2005 from Indian Institute of Information Technology, Allahabad. He obtained his Ph.D. (Computer Science & Engineering), in Feb 2010 from JIIT, Noida. He is serving as Assistant Professor at JIITU, Noida.