A NOVEL APPROACH FOR CONTEXT BASED FOCUSED SEARCH ENGINE

Synopsis of the Thesis to be submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

POOJA GUPTA

Department of Computer Science & Engineering and Information Technology
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, NOIDA, INDIA
September, 2013

A NOVEL APPROACH FOR CONTEXT BASED FOCUSED SEARCH ENGINE

The WWW (World Wide Web) is a huge repository of hyperlinked documents known as web documents, accessible through Internet [1]. It is an information sharing model that is built on the top of the Internet. WWW is a continuously growing collection of hypertext documents with its estimated indexable size of at least 50 billion web documents stored on thousands of web servers world wide [2, 3].

Owing to exponential growth of information, World Wide Web has become an important source for searching the required information. It operates on the Internet's client-server architecture. But, these web documents are not organized as books in library and no central catalogue for the same is available. Even knowing where to look for information using uniform resource locator (URL) is not a guarantee that it will be retrieved. Therefore, it is very difficult to search the desired information from such a huge collection of web documents in an efficient manner. Searching the huge WWW repository has become a challenging problem. As a solution, several information retrieval (IR) tools have been developed. These tools are divided into following three categories:

Web directories
Meta Search Engine
Search Engine

Web directories

Web directories, also called web portals or taxonomies, organize web documents into a tree-like topic hierarchy. General topics are sub-divided into more specified topics or categories. The tree-like structure of the web directories allows non-expert users to find useful information easily [4]. Some of the most popular portals on the web today are: Yahoo, Look Smart and Open Directory Project (ODP). Directories are usually human constructed by a set of experts. Web directories have two main drawbacks:

- The taxonomies are manually populated; therefore it is not possible to cover the entire web.
- Since the directory structure is populated according to the knowledge of the human constructors, different hierarchies could be built for the same number of concepts by different persons.

Meta Search Engine

A Meta search engine is a system that sends the user query to a number of search engines via a number of interface agents; collects the results from different search engines and presents them to the user. It does not maintain its own database of web documents, rather submit the search to other search engines and queries the database of other search engines. It collates the search results into one list, remove any duplicate documents retrieved from multiple sources and rank the documents according to how well they match to the user query.

The advantage of Meta search engine is that a number of different search engines can be accessed with a single query. Moreover, as the matching strategy of search engine is different from each other, there is high possibility of getting irrelevant documents in search results. MetaCrawler [5], Dogpile [6] and 37.com are examples of Meta search engines.

Search engine

Search engine are information retrieval systems that help users to find the desired information on the web. A typical web user expresses his need via a set of query terms submitted to search engine. Search engines maintain large number of web pages [7] and easily find several thousand of matches for an average query. This is done with the help of web crawler. Crawler contacts servers on the Internet and collects information in form of web documents. The downloaded documents are stored and indexed in local repository of the search engine. To resolve a user query, search engine consult its local database to produce a list of relevant web documents containing the required information. Finally, a ranked list of the URLs pointing to the relevant documents is displayed to the user. For instance, AltaVista, Google, Excite,

Bing etc. A search engine, indexes thousands of pages per day as compared to the limited manual collection of web pages in Directories.

Efficiency of all these three IR tools is evaluated in the terms of precision metric. In designing any IR system, the main objective is to improve precision so that more number of relevant documents are provided to user in return of a query response.

Thus main focus of the work presented in the thesis is to devise a mechanism that can provide more relevant documents w.r.t user query. We have designed an approach for a focused search engine that not only considers a user query context but is also capable to give high precision results in top 'k' position.

FOCUSED SEARCH ENGINE

The focused search engine concentrates on the quality of the information rather the quantity. It seeks, acquires, indexes and maintains pages on a specific set of topics that represent a narrow segment of the web. The main challenge in the focused search engine is to devise a method to decide the relevance of web pages w.r.t a specific context of query keyword. Thereafter based on the quality and desired relevance, the size of result set is reduced by removing irrelevant documents from it.
The various existing focused approaches are divided in two broad categories based on method they use to determine the relevance of web documents.

Link Structure Based Approaches

- Fish-Search System
- Shark-Search System
- Category Taxonomy
- Page Rank
- HITS
- Automatic Resource Compiler
- Context Graph Based

Query Context Based Approaches

- Using Query Context in IR

Personalized Search

Query Log Analysis

PRESY

LINK STRUCTURE BASED APPROACHES

A focused search attempts to find out relevant documents specific to a user query by exploring the link structure of the Web to evaluate relevance of a web page [8, 9]. Various types of links present in a web page are internal links, external links, transverse links and intrinsic links. The links pointed to and emerging from the web page are important resource to check the topical relevancy of a page.

Figure 1 shows the link-structure of the web represented as a directed graph with web pages as nodes and hyperlinks as the directed edges.

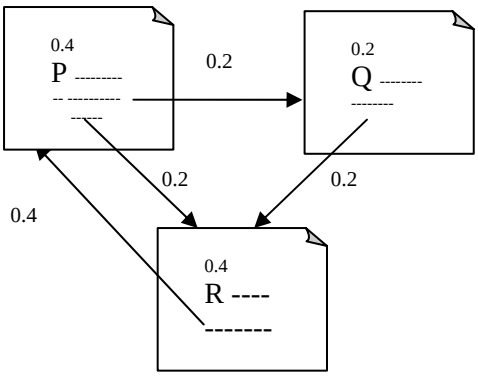

Figure 1: Link Structure of Web

As shown in Figure 1, the web pages P, Q and R are the nodes and the arrows (edges) are the hyperlinks present in them. There are two kinds of hyperlinks present in this Figure, namely in-links and out-links. The links on a web page that are pointed by other pages are called in-links or back-links. The links that are emerging from a web page to other pages are called out-links or forward-links. For Example, in the above figure, the web page R is having two in-links (pointed to it) and one out-link (pointed to P). So, P and Q are the back-links of R, R is back link to P and P is back link to Q. Further P is forward link of web page R and vice versa.

For a particular web page, once it is downloaded, its entire forward link can be known but same is not true for the back links. For computing the number of back links, all web pages pointing to it should have already been downloaded. As a result a repository is built up at the backend of every search engine that helps to compute the back link of a downloaded webpage. The various focused approaches based on the link structure of the Web are as follows:

Fish Search System

The fish search system is based on the depth first algorithm. The relevance of a document is judged based on the regular expression or occurrence of keyword in its content. If the occurrence is high the document is considered relevant and a score '1' is assigned. If the frequency of occurrence is low, the document is considered irrelevant and a score '0' or '0.5' is assigned [10]. Further, out-links from only a relevant page is explored up to a specified depth level. The links from irrelevant documents are not explored. The relevance score assigned to documents are either'1' or '0' which implies very low differentiation of the priority of pages in the list. For example, if two different documents are found to be relevant and a score '1' is assigned. Out of these two links, the link at the highest position in the list will get the priority. Whereas, link at a lower position may have more relevant out-links from it. Moreover, due to time limit the links from this page may not even get explored. Secondly, matching is done purely on the basis of occurrence of keyword or expression in the document. It is not always necessary that a document containing a word with high frequency will be relevant.

Shark Search System

The shark search system is an improvement of the fish search system. The major improvement is to judge the relevance of a web page w.r.t to a query. The similarity between anchor text around the link and the query is computed using similarity function sim(q,d) [11]. Moreover, the relevance of the parent page of the link also contributes to its relevance. An out-link from a relevant parent is considered to be relevant and then its similarity score with the query is computed. Thus, the link relevance is judged by the integrated relevance score of its parent and its similarity score w.r.t to query. Hence, as the depth increases, score of all hierarchical parents are added up to its similarity score to judge its relevance. The limitation of this approach

is that similarity is computed using anchor text only; which is a very small length text generally 2 or 3 lines about document contents.

Category Taxonomy

In this technique, the classifier learns to recognize the relevance of a web page based on the category tree document taxonomy and some seed documents. The user collects some URLs that are examples of interest such as book mark pages. These examples help a user to build the taxonomy classes and user further marks them as good [12]. He may move the document from one category to another. The classifier integrates refinement required from the user into its statistical class models. There is a distiller component which further identifies the documents containing large number of relevant resource links, called hubs. The user again marks the pages as good or changes the class. Thus, feedback from user goes back to the classifier and then to distiller. The drawback is that it requires user intervention for learning.

Page Rank

It is the most popular ranking algorithm. It judges the relevance of the web page in terms of its popularity i.e. how many popular links in Web points to it [13,14,15.16]. Thus, the quality and number of in-links to a page is major factor contributed to page rank score [17,18]. The simplest page rank computation is shown in Figure 1. This algorithm builds a probability distribution over web pages i.e. sum of scores of interlinked web pages will be one. In Figure 1, the computed page rank values of P, Q, R are 0.4, 0.2 and 0.4. The Chattamvelli [19] has discussed about various generalization to the original page rank algorithm such as NoRPRA (Noise Removed Page Rank Algorithm), APRA (Alpha Page Rank Algorithm), WePRA (Weighted Page Rank Algorithm), FiPRA (Filtered Page Rank Algorithm) and HyPRA (Hybrid Page Rank Algorithm). The page rank is purely link structure based algorithm. Context of query keywords is not considered while computing the relevance. Another problem with this approach is that every new page added to the Web with less number of links to other pages is given a low score value and get displayed at low position in results list whereas it may be more relevant to context of query keywords. Moreover, the number of in-links to a page can be easily manipulated; causing irrelevant documents to appear at top position in result list.

HITS

The HITS is another ranking algorithm that judge the relevance of a web page based on two heuristics [20, 21]. First heuristics is authority pages which are the pages that are relevant, popular and focused to a particular query. The in-links to these authority pages are considered as second heuristic called hub pages. Hub pages (back-link pages) are web pages that contain useful links to relevant pages and also links to many authorities. The hub score and the authority score of a web page together determine the relevance of a web page.

Automatic Resource complier (ARC)

This approach gathers the web documents using HITS like algorithm. It automatically compiles a source list on any topic [22]; modify weighted authority and hub score. It uses anchor window on either side of the href to check the similarity. The weights are assigned depending upon the similarity of anchor text around the link. An iterative process, based on modified weights computes the hubs and authority pages. Thus, relevance of a link is judged on the basis of similarity between anchor text windows, present in the parent page. Its main drawback is use of very limited text to judge the relevance against a query which may or may not contain the exact topical information of the link.

Context Graph Based

The relevance of a web page is judged based on the learning in the form of context graph. It builds a context graph from initial downloaded web page on a topic and its 40 highest positioned in-links. This graph represents the hierarchical relationship in web pages. The target page is placed at the bottom layer. Their immediate parents (in-link) are placed at one level higher layer and so on. This graph is used to train the classifier [23]. Context graph constructed for various topics are then merged to a single graph. Each layer in the graph is considered as a category class. The context graph provides links to other related topic pages and provides back-ward search facility. The relevancy of new page is judged based on its TF-IDF probability score against each category class and is added to the corresponding class. Thus this approach considers the relevance based on TF-IDF similarity score computation w.r.t to category classes. Moreover, the top 40 in-links (back-links) to a web page are considered equally important to be place in the repository.

QUERY CONTEXT BASED APPROACHES

The aim of the context based focused approaches is to get the context of the web pages, in other words these approaches try to find the topic to which the page suits the best. Some approaches try to get the exact context of the user query keywords, so as to find more related web pages. Some of the context based approaches are discussed in detail in the following sub sections.

Personalized Search

The aim of the focused search is to display more relevant pages in response to a user query. However, query keywords passed by the user provide less information to specify the user exact need. The personalized search system identifies the user interest in terms of implicit feedback from user. By capturing user's recent activities like previous search queries, visited pages, recently read/created mails; a user current search need is identified [24, 25, 26]. The exact context of query keywords is specified based on the user recent behaviour. For example, if a computer programmer has passed a query 'java' the identified context will be in the sense of 'java as a programming language'. This sometime results in irrelevant documents as user interest may change from time to time. So, it is inappropriate to decide query context based on recent search history.

Using Query Context in IR

This technique integrates the query-specific context and context within query to get the exact context. A query related profile is maintained in spite of user related general profile [27]. This includes the user domain of interest. The context within query implies to add the more query words in the query vector by introducing strict condition in relations between terms. The co-occurrence relation between the initial query terms is judged. Using the global resource information (from Global Dictionaries) more related words are added to the query vector depending upon probability of co-occurrence of this word with query terms. The new word is added to the query vector if the probability of its co-occurrence with the initial query terms is greater than a threshold value. Thus by adding more query words disambiguation between terms is removed. This model requires at least two query terms in user's initial query.

Query Log Analysis

In query log analysis method, annotations of queries, sessions and actions are analyzed from query logs. The proposed approach in [28] considers user's query history and click through document history within a particular session as implicit feedback. Each query is assigned a unique Id and the results displayed are first transferred to a proxy-server that records the links with query Id. The relevance of a new resultant document is judged based on the recorded history and more relevant documents are displayed to the user. However, the same query in future may be passed by the user but in different context. For example, the 'jaguar' referred to as animal in earlier session may change to automobile context in current session. As a solution, the idea of considering the current session history or activities as implicit feedback [29] was proposed. The implicit feedback in terms of query history and click through history within the same session is used to get the context. Thus, the recent or current history of the user is used to check the relevancy of the matched documents in response to a user query.

PRESY

A profile based reformulation system is designed to get the context in which user's profile has been used. The approach is based on the static and dynamic context which uses the context from the user's profile to automatically reformulate the initial query by appending more terms from the context of current search [30]. The performance of the mechanism results in more relevant documents only if the user profile is constructed well.

From the critical analysis of the available literature following conclusions are drawn:

1. All existing link-structure based approaches like page-rank judge the relevance of a web page by using its link structure and dose not consider the context of query keywords. More the number of popular links pointing to a page more important that page is considered and thus assigned a higher score. In other words, the number of popular in-links and relevant out-links is the criterion to judge the relevance of a web page. There is no relation established between the popularity of a web page with context of query keyword.

2. Other approaches have used simple TF-IDF over vector space model, regular expression matching, similarity between anchor text and query as various criteria to judge the relevancy of web page w.r.t. to a user query. These approaches are based simply on occurrence of query keywords in web documents. No significance is given to actual location in the document where these query terms occur.

3. Number of back-links pointing to a web page is used as the criterion to judge the popularity of a web page. All back links are considered equally relevant and none of the approaches have actually ranked them on context of query keywords. Most of the approaches have not included the back link pages in the final result to improve precision.

4. Existing approaches have identified the context of query keywords from user search history, user profile, query logs, user behavior, and documents stored at desktop etc; but have failed to resolve the ambiguity in query keywords itself arising due to polysemy or hyponyms. For Example, the search keyword ‘java’ leads to various contextual senses such as ‘java programming language’, ‘java coffee beans’, ‘java island’ and ‘jaguar’ leads to ‘jaguar automobiles’, ‘jaguar animal’, ‘jaguar fittings’ etc. Thus, there is a need to identify different contextual meaning of words and to remove ambiguity.

5. Generally, current commercial engines provide an interface where a user types in query keywords. They do not capture different context of query keywords as a result they return thousands of matched documents in response to a query, however, only a fraction of results are valuable to the user. This size of information provided by a search engine is too large to go through. This leads to the **problem of information overkill**. This problem further aggravates in case of inexperienced users, trying to search the information from the web. According to the famous “8 second rule” [31, 32] such users look at first few results and tend to turn away. It is evident that these search engines lack in the ability to figure out the exact context of a user query.

Thus, objective of the thesis work is to propose & implement an approach for
A Context Based Focused Search Engine that:

1. Considers the various contextual senses of the keywords present in user query and web pages.
2. Evaluates the relevance of web pages based on various contextual senses of keywords present in them
3. Further, enhances search results by including selected back-links to web pages based on contextual senses

And hence the title: "A Novel approach for Context based Focused Search Engine"

In order to design and implement such a context based focused search engine, following issues are identified and contributions are made to address the issues.

Issue 1: Some keywords have multiple contextual senses or meanings known as 'polysemy'. For example, in English 'mouse' is a pointing device in computing and rodent elsewhere. Polysemy can also be categorized as noun-polysemy and verb-polysemy etc. the keyword 'mouse' is an example of noun-polysemy. There are many words which, when used as noun, verb etc. lead to different meanings. For example, 'fly' if used as noun refers to an 'insect' and if used as verb refers to 'act of moving in the air'. These different meanings are also called different contextual senses of words.

Whenever any such keyword is given to a search engine, it fails to capture the actual context of the keyword that a user desire before processing the query. As a result documents are evaluated for multiple contextual senses, and consequently large numbers of documents are returned in response to a user query having the same keyword but in different sense. A user has to browse the results to find documents matching his desired context.

In order to address **issue 1**, Word Net dictionary [33, 34, 35] has been integrated to design context based focused search engine to get various contextual senses of keywords. Our proposed solution presents to a user a collection of different contextual senses and allows him to select desired contextual sense. An interface with the Word

Net dictionary has been implemented in Java as front end and oracle 10g express as backend. This module takes keywords as input and return < meaning, definition> pairs. This thesis has contributed to address the problem of Information overkill at the query interface level.

Issue 2: A web page can be evaluated with respect to different contextual senses for instance a mouse as a rodent and mouse as computing device. But mostly, a page is suited to one of the senses. There is a need to devise a mechanism that can score a web page and determine its relevancy w.r.t. to a contextual sense.

In order to address **issue 2**, the thesis has proposed a context based relevance evaluation mechanism to find the relevance score of web document in terms of its contextual score. Backend repository of documents can be built up for search engine in which documents are ranked on the basis of contextual score. The proposed contextual sense based relevance evaluation will help the search engine to identify the context of the document. This contribution will result in placing contextually more relevant documents at top positions in search results. The mechanism, if integrated with existing link based technique will help the search engine to display more relevant as well as popular web documents to the user at top positions in result list.

Relevance evaluation is applied at query processing end to rank the web documents. It can also be applied at crawling end, where crawler can judge the relevance before downloading documents. This will help in controlling size of backend repository.

Issue 3: Back links are a potential source of information in a given topical area [36, 37]. Search results precision can be improved by replacing some of the irrelevant or less relevant pages with more relevant back link pages. However, all back-link pages may not be equally relevant to a user query. There is a need filter back-link pages based on the relevance before including them in results.

In order to address **issue 3,** in this thesis a back link extraction algorithm has been proposed and implemented that not only extract the back links of downloaded web pages but it also computes the relevance of the extracted back links taking the associated web page as the base. The analysis of the computed results is also done to

filter back links based on their contextual score. The relevance evaluation of back-link and consideration of only relevant back-links in results will enable the search engine to store the more number of relevant related documents for a contextual sense. This will increase the overall precision results of search engine.

Issue 4: Indexing play a vital role in search engine as a database index optimizes time and computing requirement to answer for a search query. An index in most of the commercial search engines generally contains all different terms in each document sorted in alphabetical order and an associated list of documents that contain the term. This kind of inverted index allow a search engine to carry a search by first finding the match in index and then follow the respective list of documents, irrespective to the full scan of each document. A current index structure does not store contextual information in terms of contextual score. There is a need to design an index structure to incorporate contextual information with the keywords to make search process faster.

In order to address **issue 4,** in this thesis we have designed and implemented a context based indexing scheme for web documents. The proposed tree based inverted index can store various contextual senses of keywords and a list of document pointers, that have the similar contextual sense, associated with that keyword.

Issue 5: In order to design a context based focused search engine various components like relevance evaluator, back link extractor and indexer needs to be integrated together in a unified architecture.

In order to address **issue 5,** this thesis has proposed a complete architecture for a novel context based focused search engine which integrates all the components that resulted after addressing all four issues.

This thesis has also resulted in design and implementation of a prototype for a context based focused search engine.

The prototype is evaluated on 50 different query keywords having approx. 115 different contextual senses and can be scaled further. For each contextual sense, 20 web page URLs and 80 back-link page URLs are downloaded (4 back-link URLs for

each web page URL) and a backend repository is built. Approximately, total 11,000 URLs have been evaluated using context based relevance evaluation mechanism. The analysis of the results shows that not all the back-links to a web page are equally important. The details of results are included in the thesis. Search results have shown improvement by including the back-link URLs. More number of relevant documents is displayed to the user earlier at top position. The average score get improved from 0.31 to 0.41, if back-links are considered.

The ranking order of the proposed mechanism has been compared with the ranking order of the same documents using page rank algorithm values. It has been observed that proposed technique placed the contextually more relevant document at top positions. The average precision has improved from 0.65 (for page rank score) to 0.82 (proposed relevance score). Thus, number and quality of relevant documents displayed to the user have increased.

The contextual sense based user interface designed has been compared with the Google current interface. Google provides an auto-compete facility in search text box that automatically guides the user with various search options. It has been found that Google generally covers the synonyms of the query words but fails to filter results on polysemy. The solution proposed in this thesis will cover the polysemy meanings of the query keywords. Further, the results displayed by the Google for the same contextual sense are compared with the results displayed using proposed context based focused search. The analyses of results have established that similar documents are displayed at different positions; and the proposed ranking mechanism displays the contextually more related document earlier at top positions.

The limitation of the work done in this thesis is that it is valid for the textual documents only and documents containing images and video are not considered. Moreover, the work done is dependent on Word Net dictionary for various contextual senses. The technique failed in case of proper noun where Word Net does not provide contextual senses.

Organisation of the thesis – The thesis is organized in seven chapters. A brief review of each is as follows

Chapter 1: This chapter covers the introduction about World Wide Web (WWW), internet and search engine and web crawlers. The evolution of WWW and various search tools has been discussed.

Chapter 2: This chapter focus on basic concepts of various information retrieval tools. It describes the architecture of a general search engine. It provides detailed review of existing techniques in the field of focused search, the link structure of the web and the various existing ranking algorithms. A study on user search trends is also provided. Based on the literature review, major challenges and limitations in existing approaches are identified.

Chapter 3: In this chapter, a novel architecture for context based search engine has been proposed. The proposed architecture uses web pages as well as their back links URLs. The architecture consists of three main layers. Bottom layer collects the downloaded web documents and back links in local repository and maintains the index of local repository. The middle layer computes the relevance score of each matched result searched by query processor and then pass the ranked list to upper layer, to display to the user. The detailed working and algorithms of bottom layer components has been provided.

Chapter 4: This chapter elaborates mechanism for back link extraction and its relevance evaluation. Algorithm design, its implementation and results are also discussed.

Chapter 5: This chapter has proposed a ranking mechanism based on the contextual senses. This chapter shows how the contextual characteristics can be used to evaluate the relevance of the web documents. The performance evaluation of proposed algorithm is discussed using standard metrics Precision and Recall.

Chapter 6: In this chapter, a context based indexing of web documents has been proposed and discussed. It presents a modified index structure to store the contextual characteristics of keywords with the inverted index of keywords.

Chapter 7: This chapter presents design, implementation and testing results of prototype built for a context focused search. Implementation of the user interface is discussed in detail under this chapter.

REFERENCES

[1] Berners-Lee, Tim, "*The World Wide Web: Past, Present and Future*", MIT USA, Aug 1996

[2] "Internet World Stats. Worldwide Internet Users", available at: http://www.internetworldstats.com (accessed on Feb 21, 2013)

[3] Maurice de Kunder, "Size of the World Wide Web", available at: http://www.worldwidewebsize.com (accessed on Feb 21, 2013)

[4] Gerstel, O., et al "*Reducing human interactions in Web directory searches*", ACM Transactions on Information Systems, Vol. 25, No. 4, Article 20, July 2007.

[5] http://www.metacrawler.com

[6] http://www.dogpile.com

[7] Mike Burner, "Crawling towards eternity: Building an archive of the World Wide Web", Web Techniques Magazine, 2(5), May1998.

[8] Q. Tan, P. Mitra, C. Lee Giles, 'Designing Clustering-Based Web Crawling Policies for Search Engine Crawlers', Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, New York 2007, pp. 535-544.

[9] C. Benincasa, A. Calden, E. Hanlon, M. Kindzerske, K. Law, E. Lam,J Rhoades, I. Roy,M. atz, E. Valentine and N. Whitaker, "Page Rank Algorithm", 2006, http://www.math.umass.edu/~law/Research /PageRank/Google.pdf.

[10] P. De Bra , G. –J. Houben , Y. Kornatzky, R. Post, "Information Retrieval in Distributed Hypertexts" , Proc. of RIAO' 94, Intelligent Multimedia , Information Retrieval systems and Management, New York, 1994

[11] M Hersovici, M. Jacovi, Y. Maarek , D. Pelleg, N. Shtalhein , "The shark search algorithm –an application : tailored web site mapping", Computer networks and ISDN Systems, Special issue on 7th WWW conference, Australia, 30(127), 1998.

[12] S. Chakrabarti, M. Van Den Berg, V. Dom, "Focused crawling: An new approach to topic specific web resource discovery", Proc. of 8th international WWW conference, Toronto, Canada, May-1999

[13] Monika R. Henzinger, " Hyperlink Analysis for the Web", IEEE Internet Computing, 2001
URL:http://facweb.cs.depaul.edu/mobasher/classes/csc575/papers/hyperlink.pdf

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

[15] Phil Craven, Google's Page Rank Explained, Copyright Web Workshop

[16] S. Brin & L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Seventh Int'l World Wide Web Conference, 1998.
[17] C. Ridings, M. Shishigin, "Page Rank Uncovered", Technical Report, September, 2002, http://www.voelspriet2.nl/ PageRank.pdf.

[18] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry, "The PageRank Citation Ranking: Bringing Order to the Web", 1999, Technical Report, Stanford InfoLab.

[19] Chattamvelli, Rajan, "Some generalizations of the PageRank metric", National conference on current trends in advanced computing, CTAC'10, Banglore, April 2010, 172-175.

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

[21] J Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of ACM, Vol 46, no 5, September 1999, pp. 604-632.

[22] S. Chakrabarti, B. Dom, D. Gibson, J. Klienbrg, P. Raghvan, S. Rajagoplan, "Automatic Resource Compilation by Analysing Hyperlink Structure and Associated Text", 7th World Wide Web Conference, 1998.

[23] Michelangelo Diligenti, Frans Coetzee , Steve Lawrence, C. Lee Giles , and Marco Gori , " focused crawling using context graph", VLDB '00 : proc. of the 26th international conference on very large data bases, san francissco, CA, USA, Morgan KUFMann publishers inc. pp . 527-534, 2000

[24] J. Teevan, S. T. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities", In SIGIR, 2005, pp. 449-456

[25] Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T., "Personalized Search", *Communications of ACM*, 45: pp. 50-55, 2002.

[26] Belkin, N., G. Muresan and X. Zhang, "Using user's context for IR personalization. Proceeding of the ACM/SIGIR Workshop on Information Retrieval in Context, July 25-29, 2004, ACM Press, Sheffield, UK., pp: 1-3.

[27] Jing Bai, Jian-Yun Nie, Hugues Bouchard, Guihong Cao, "Using Query Contexts in Information Retrieval**",** *SIGIR'07*, July 23–27, 2007, Amsterdam, Netherlands.

[28] Sergio Duarte Torres, Djoerd Hiemstra, Pavel Serdyukov, "Query log analysis in the context of Information Retrieval for children", ACM, *SIGIR'10*, July 19–23, 2010, Geneva, Switzerland.

[29] X. Shen, B. Tan, and C. Zhai, "Context-sensitive information retrieval using implicit feedback", In SIGIR, 2005 pp. 43-50.

[30] Abdelkrim Bouramoul, Mohamed-Khireddine Kholladi and Bich-Lien Doan, "PRESY: A Context Based Query Reformulation Tool for Information Retrieval on the Web", Journal of Computer Science 6 (4): 470-477, 2010, ISSN 1549-3636

[31] Zona Research, 2001, The need for speed II. *Zona Market Bulletin*, **5**, April 2001. Available at: http://www.keynote.com/downloads/Zona_Need_For_Speed.pdf.

[32] Toronto Web Marketing, SEO / Web Marketing Articles: 8 Seconds Rule – Making your web site stickier. Available at: http://www.allanpollett.com/8-seconds-rule.html

[33] Ingo Feinerer and Kurt Hornik, wordnet: WordNet Interface. R package version 0.1-8, 2011, http://CRAN.R-project.org/package=wordnet

[34] Mike Wallace, Jawbone Java WordNet API, 2007, http://mfwallace.googlepages.com/jawbone.html

[35] Christiane Fellbaum, WordNet: An Electronic Lexical Database. Bradford Books, 1998.

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

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

List of Author's Publications:

1. P. Gupta, Sandeep K Singh, Divakar Yadav, A K Sharma, "An Improved Approach to Rank Web Documents", JIPS, Korea, Vol. 9, No. 2, pp 217-236, June-2013, **Indexed at dblp, Scopus**

2. P. Gupta, A K Sharma, Divakar Yadav, "A Novel Technique for Back-Link Extraction and Relevance Evaluation", IJCSIT, Vol. 3, No. 3,

June 2011, pp-227-238 (**Indexed at DocStoc, PubZone, DOAJ, Inspec, EBSCOS etc.)**

3. Pooja Gupta, A K Sharma, J P Gupta, "A Novel Framework for Context based Distributed Focused Crawler", IJCCT, Vol. 1, No 1, 2009, pp-14-26 **(Copyright Inderscience Enterprise Ltd.)**

4. P Gupta, A K Sharma, J P Gupta, "A Review of Indexing Techniques", M R International Journal of Computer and Technology, 2009

5. P Gupta, A K Sharma, S K Singh, "A Novel Context based Indexing of Web Documents", IEEE International Conference at Rajkot, May 11-13, 2012, pp-448-452 **(Indexed at IEEE-Xplore, Scopus)**

6. P Gupta, "Context based Relevance Evaluation of Web Documents", IC3-2012, CCIS-306, pp-201-212 **(Indexed at Springer, Scopus, dblp)**

Appendix I

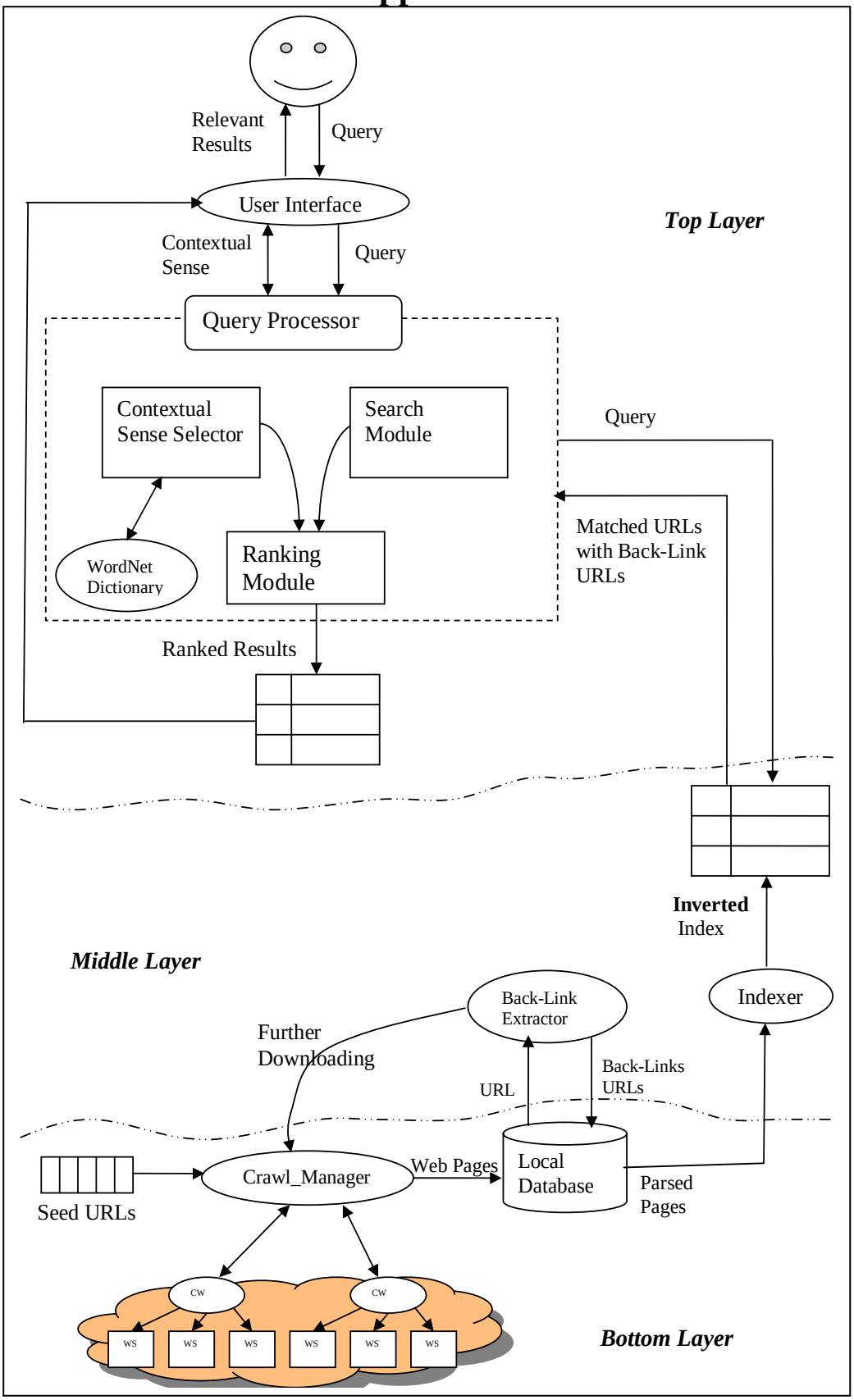

Figure 2: High Level Architecture

Supervisors

Dr. Sandeep K Singh	Dr.Divakar Yadav	Prof. A.K.Sharma
Asst. Professor	Asst. Professor	Prof & Dean (PG & Research)
JIIT, Noida	JIIT, Noida	BSA, Faridabad