pISSN 1976-913X
eISSN 2092-805X

J Inf Process Syst, Vol.9, No.2, June 2013 http://dx.doi.org/10.3745/JIPS.2013.9.2.217

An Improved Approach to Ranking Web Documents

Pooja Gupta*, Sandeep K. Singh**, Divakar Yadav** and A. K. Sharma***

Abstract—Ranking thousands of web documents so that they are matched in response to a user query is really a challenging task. For this purpose, search engines use different ranking mechanisms on apparently related resultant web documents to decide the order in which documents should be displayed. Existing ranking mechanisms decide on the order of a web page based on the amount and popularity of the links pointed to and emerging from it. Sometime search engines result in placing less relevant documents in the top positions in response to a user query. There is a strong need to improve the ranking strategy. In this paper, a novel ranking mechanism is being proposed to rank the web documents that consider both the HTML structure of a page and the contextual senses of keywords that are present within it and its back-links. The approach has been tested on data sets of URLs and on their back-links in relation to different topics. The experimental result shows that the overall search results, in response to user queries, are improved. The ordering of the links that have been obtained is compared with the ordering that has been done by using the page rank score. The results obtained thereafter shows that the proposed mechanism contextually puts more related web pages in the top order, as compared to the page rank score.

Keywords—Ranking, Ordering, WWW, Information Retrieval, Contextual Relevance, Contextual Sense, Web Documents

1. INTRODUCTION

WWW is a huge repository where the information is stored in the form of hyperlinked web documents. It is where each page on it can be reached through other pages by following the hyperlinked path that is present at them. Owing to the dynamic nature of the web, the repository is continuously growing. Therefore, the search engine's task has become more challenging in the sense that it has to filter out the desired content from this huge repository in response to a user query.

The search engine works in two phases. In the first phase it parses the user queries and then the matched contents from the repository are extracted accordingly. Generally, the matched results are thousands in number, and users are not interested in such a large result set. It is a challenging task for a search engine to decide the order in which the result should be displayed to the end users. So, in the second phase, the search engine uses a relevance rank combined with some other measures and heuristics (popularity of the page, TF/IDF, etc.) to rank the documents so

Manuscript received October 29, 2012; first revision October 13, 2012; accepted February 14, 2013.
Corresponding Author: Pooja Gupta
* Dept. of Computer Science and Engineering, I.P. University, Delhi, India (poojaguptamait@gmail.com)
** Dept. of Com Science and Engineering, JIIT (deemed to be Univ.), Noida, India ({sandeepk.singh, divakar.yadav}@jiit.ac.in)
*** Dept. of Comp Science and Engineering, Y.M.C.A. Univ, Faridabad, India (ashokkale2@rediffmail.com)

that the most relevant pages can be placed at the top in the ranking order.

Generally, the search engine uses the link-structure of the web [1, 2] and its various properties to assign some numeric weight to the page. This is called ranking. The link-structure of the web can be represented as a directed graph with web pages as nodes and hyperlinks as the directed edges.

For example, in Fig. 1 the web pages P, Q, and R are the nodes and the arrows (edges) are the hyperlinks that are present in them. There are two types of hyperlinks, in-links and out-links. The links that are pointed at by another page are called in-links or back-links. The links that are emerging from the page to other pages are called out-links and forward-links. For example, in Fig. 1 the web page R has two in-links (pointed to it) and one out-link (pointed to P). So, P and Q are the back-links of R. For a particular web page, once it is downloaded all of its forward-links can be known, but the same is not true for the back-links. Generally, a page is considered more important if it is highly linked with other pages. These links play an important role in ranking the web pages. The page rank [3] also depends on the number and quality of back-links. Chattamvelli [26] has discussed various generalizations for the Original Page Rank Algorithm (OPRA) such as NoRPRA, APRA, WePRA, FiPRA and HyPRA. However, the simplest mechanism applied for the page rank computation [25] can be represented by the following expression:

$$R(u) = c\sum_{v} \frac{R(v)}{N_v} \ni v \in B_u$$

Where,

u: the web page

R(u): the page rank of 'u'

B_u: back-links

Nu: |Fu| the number of links from u

c : the factor used for normalization

The Page Ranks form a probability distribution over web pages, so the sum of all web pages'

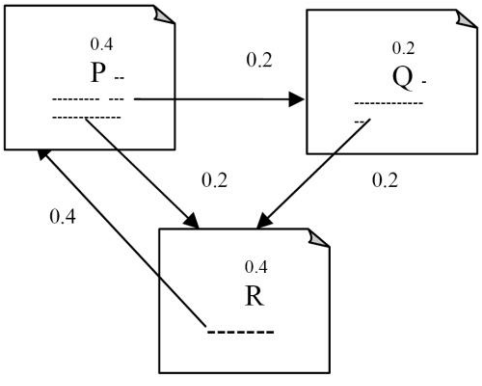


Fig. 1. Link Structure of the Web

Page Ranks should be one. The page rank values for page P, Q, and R are computed using the above expression and are shown in Fig. 1. For the purpose experimentation in this body of work, we have used the simple Page Rank Score.

Similarly, the HITS [5] algorithm ranks the page based on the link structure where the terminologies used are the Hubs and Authority pages. HITS considers the back-links (in-links) as the hub pages and the pages to which these in-links points, as authority pages. This is where the authority pages are the pages that are relevant, popular, and that are particularly focused on the query. Hub pages (back-links) are those that contain useful links to relevant pages and also links to many authorities.

None of the existing ranking methods consider the contextual senses (polysemy) of the keywords that are present in the web page while computing the rank of the web documents against user queries. A keyword may have various meanings depending upon its usage. This is called the contextual sense. For example, the keyword 'Server' can be used as waiter, court game, computer science host, and as a utensil.

Furthermore, it has been observed that existing mechanisms use the back-links count to rank a web page and to consider all the back-links to a web page. The information in a particular web page might be partial in nature and sometimes does not fulfill the user requirement completely. As a solution, Soumen [21], has found that if the backward link (back-link) information is provided to the user, it will enhance the knowledge discovery process. The back-links to referral parent pages provide information about the related web resources.

In our previous work [4], we have analyzed the relevance of back-links to a web page. It has been observed that not all of the back-links to a web page are equally important. The back-links are filtered out based on the contextual senses of the keywords. In this paper, to rank a web page the contextual senses of the keywords that are present in them are used and only the contextually related back-links are considered as a good source of topical information. The proposed mechanism computes the probability measure for each contextual sense that is related to the keywords that are present in the web page to assign a rank value. Similarly, the rank values for back-links are computed. All the pages thus obtained (web pages and their back-links) and are then ordered according to the computed rank value. The experimental result analysis shows that when using this technique various highly ranked, relevant back-links are displayed in the top positions, which results in replenishing the user with more related information.

2. RELATED WORK

Generally the search engines index the retrieved web documents based on the popularity of the pages on the web. However, the popularity of the web page may change over time as suggested by Klienberg [5], wherein it distinguishes two types of pages as being 'hubs' and 'authorities'. Hubs are the connection points of important pages that point to important authorities. Authority is a page that is pointed to from many important hubs. The hub-score H(x) and authority-score A(x) that are associated with page x are recursively updated. Another popular method of ranking the documents, which was introduced by Brin and Page [6], is called the page rank algorithm (mentioned in Section 1). It is a relative measure that is used for ranking the importance of a web page with respect to the other web pages that link to it. Some search engines to relevance rank retrieved documents on a high-to-low priority order use it. Researchers have published

various extensions of the Page Rank algorithm. The Trust Rank [22] is an extension of the Page Rank. It aims to reduce the side effects of spam and unwanted web pages, so as to improve the quality of the Page Rank.

The most popular HITS and Page Rank mechanism work mainly on link structures. Some researchers have computed the rank of a web page based on their content analysis. One such mechanism was proposed by Emil Gatial [7]. It computes the page relevance based on the keywords that are present in the web page by comparing keywords that match in relevant or irrelevant pages and it assigns a weight based on that. The sum of weights of all keywords divided by the number of keywords in the page is considered to be a page relevance score, which is considered to be the rank. Z. Liu [8] has proposed a method based on domain ontology and Formal Concept Analysis (FCA). It first constructs a core similarity graph (CSG) using WordNet ontology and concept relatedness and then constructs the Similarity Concept Context Graph (SCCG) based on CSG and FCA. The crawling strategy measures the expected relevancy of a page to a given topic using SCCG. It also determines which URL should be crawled first.

The page rank algorithm [9] ignores which page the outgoing link points to. It considers all of the outgoing links that are present in a web page as being equally likely. All outgoing links are weighted in equal proportion and give equal importance to each of the pointing pages. In this algorithm, if two web pages have an equal page rank, it implies that both the web pages are equally popular. However, it does not indicate what the content similarity of those pages is.

It has been observed from the existing work that the hubs and authorities (HITS) algorithm may not distinguish between the hubs and authorities when there is a cycle present in the graph. The hubs and authority score computed by HITS are barely applicable for similar queries and in similar areas. While the web pages often have various domain contents, the hub and authority score cannot signify the general relevancy of the web pages.

M. Persin [10] has explained the need for the ranking technique to find the answer to a query in the document collection stage. He has used the cosine measure to find the similarity of a document and a query. Each term is assigned a weight according the TF-IDF method. Each non-zero value for a similarity measure is divided by the weight of the document and the top ‘k’ results are displayed to the user.

A critical review of the available literature shows that none of the existing works have used the contextual senses of the keywords while computing the relevance of a link or a web page.

In this paper, a ranking mechanism is being proposed that ranks the web pages and their back-links based on the different contextual senses of keywords and orders them according to the computed rank. The proposed mechanism is quite different from the other techniques as it assigns different scores to each back-link, so not all the back-links are considered equally. It assigns weight to each relevant term that is present in a web page and in its back-link as per its presence in various HTML tags. It computes the rank depending upon the conditional probability of the contextual sense (CSense) to occur in the current web page. Finally, the top ‘k’ web pages and back-links are displayed to the user.

3. A Framework for Ranking Web Documents

The proposed ranking mechanism uses an amalgam of contextual senses of the keywords present within pages and their back-links to rank the documents. The page and its back-links are

thereafter arranged based on the rank obtained. The proposed method works in the following two phases: in the first phase it computes the contextual senses of the keywords within a web page and its back-links, and in the subsequent phase, it ranks the page and its back-links. Thus, the two phased ranking system helps to identify the best suited web pages for a given user query. The proposed architecture for the ranking mechanism is as shown in Fig. 2.

The architecture consists of four main components: the crawler, back-link extractor, indexer, and query processor.

The crawler continuously downloads the web pages and stores them in the repository. The back-link extractor extracts the list of all the back-links of a URL and accordingly updates the repository. For this purpose, the *Back-link* extractor, which was developed by the author in [11], is being used. It extracts all the back-links of the web page by using the recursive search of links in the repository of downloaded pages. Then, the indexer creates an inverted index of the documents in the repository.

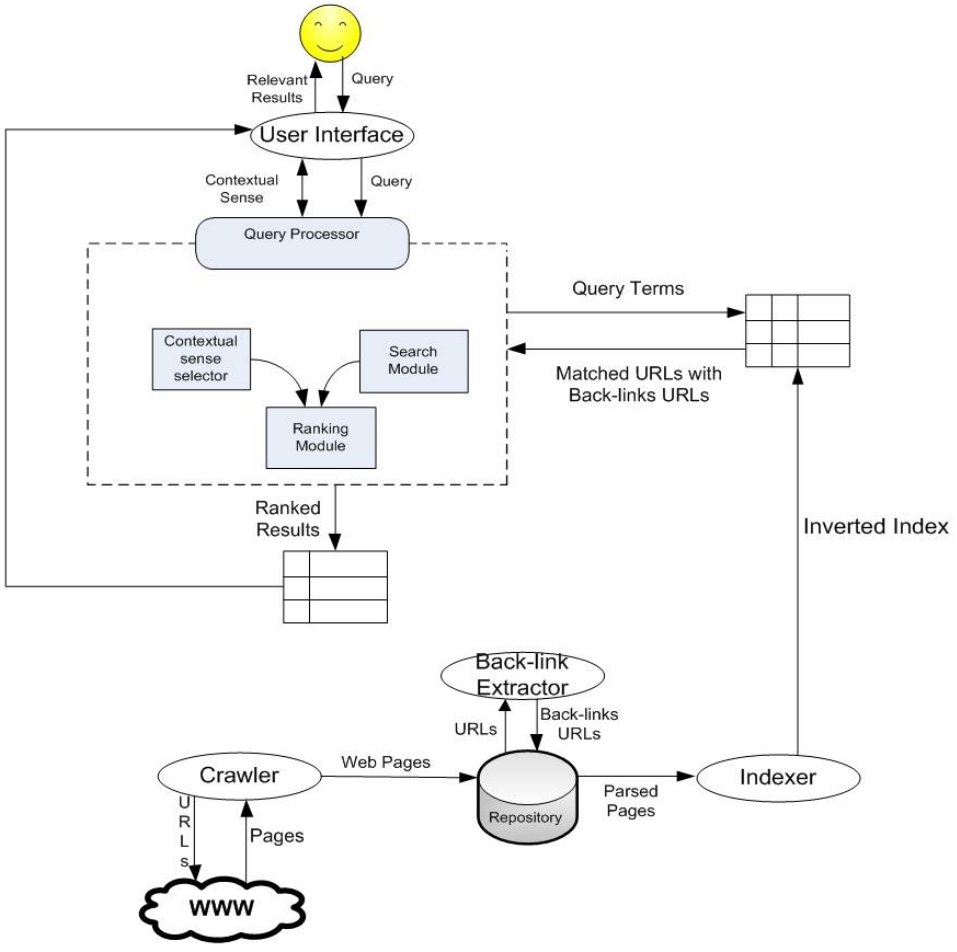


Fig. 2. Architecture Framework

The query processor is the component that matches the query to the index of documents. The documents that are matched with the query requirement are assigned a similarity score based on a ranking algorithm. Then, based on these computed ranks/scores it presents an ordered list of matched documents to the user.

The query processor in the proposed architecture is composed of three sub-modules, which are called: the *Contextual sense selector* module, the Ranking module, and the Search module. The *Contextual sense selector* module gets the user query from the user interface and contacts the contextual sense database to get the contextual senses of the query keyword and asks the user again to select a particular sense. The Search module searches the inverted index to get the list of ordered pairs of URLs and back-links that are matched to the user query. The ranking module computes the rank using the proposed mechanism. The ranking algorithm presented in this paper computes the rank of the web page and its back-links (retrieved by the search module) based on the contextual senses of the keywords. The highest ranked web pages and back links for a user query are presented in the selected sense to the user as the topmost results.

3.1 The Ranking Module

The *ranking module* is further composed of the keyword extractor, the contextual sense extractor, and the *Rank calculator* components. Fig. 3 shows the computation flow of the *ranking module*.

The *ranking module* takes the list of retrieved web pages and their back-links as input from the search module. The *keyword extractor* extracts the list of keywords, their relative frequency

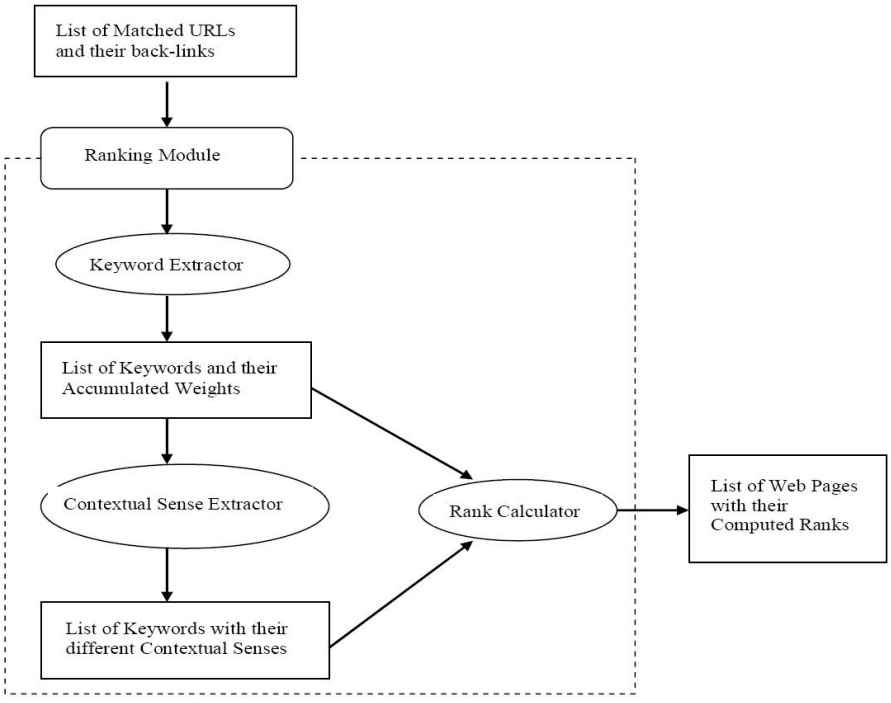


Fig. 3. The Ranking Module

at various tags, and computes the relative accumulated weight of each keyword. Furthermore, the *contextual sense extractor* extracts the contextual senses of the extracted keywords. Finally, the r*ank calculator* calculates the rank of the web page and their back-links based on the accumulated weights and contextual senses of the keywords. In this work, for the purpose of result analysis the ranking of documents is done based on different contextual senses of the first 5 highest weighted keywords.

The functions and characteristics of all components of ranking module are discussed in detail below.

3.1.1 Keyword Extractor Module

An HTML document is a structured hypertext document. A keyword may be present throughout anywhere in the HTML document with varying frequency and may also be associated with various tags. The text segments that are marked by various HTML tags have specific meaning. From the work of [13], it is known that every HTML tag has a weight that is relative to its importance in the HTML document.

The keyword extractor module designed in this work extracts the keywords that are present in a web document and counts their relative frequencies in various HTML tags and assigns weights to the HTML tags. To calculate the weight of a keyword, we assigned different weights to various HTML tags as per their relative importance as discussed in [13, 14]. Subsequently, it computes the accumulated weight of all the keywords as shown below:

$$AW_{k_i} = k_{it} * w_t + k_{ib} * w_b + k_{ih} * w_h + k_{il} * w_l \quad (1)$$

Where,

AW_{ki} : the accumulated weight of keyword k_i

k_{it}: the frequency of occurrence of keyword k_i in the 'title' tag

k_{ib}: the frequency of occurrence of keyword k_i in the 'body' tag

k_{ih}: the frequency of occurrence of keyword k_i in the 'head' tag

k_{il}: the frequency of occurrence of keyword k_i in the 'link' tag

w_t: the weight assigned to the 'title' tag (.12)

w_b: the weight assigned to the 'body' tag (.32)

w_h: the weight assigned to the 'head' tag (.26)

w_l: the weight assigned to the 'link' tag (.3)

Numerical values inside parentheses indicate the weight that is assigned to different HTML tags.

Table 1. The occurrence of keywords and their corresponding accumulated weights

Keyword	Title	Head	Link	Body	Accumulated weight
Mouse	0	2	31	146	56.54
Mice	0	6	8	61	23.48
Button	0	0	2	32	10.84
Computer	0	0	9	24	10.38
Mouse	1	2	20	10	9.84

For instance, in the following URL

URL: http:\\en.wikipedia.org\wiki\mouse_(computing)

The top 5 keywords and their accumulated weight according to their occurrence and frequency in various tags are given in Table 1 above.

3.1.2 The Contextual Senses Extractor Module

A keyword can have multiple senses. Thus, it's a challenging task to identify the different contextual senses of a keyword. "The different meanings of keywords in a different context," is called 'Polysemy.' For example in English a 'mouse' means a pointing device for computing and means a 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 "the act of moving in the air." These different meanings are also called different contextual senses of words. Our proposed approach is able to handle the polysemy property of keywords. It extracts the various contextual meanings of a keyword and presents these meanings to a user as a collection of different senses.

In the proposed work, for the purpose of getting the various contextual senses of keywords, we have used the WordNet dictionary [15-17] and thus the contextual senses of definitions obtained from WordNet are used to find the contextual sense of a web document. This module takes keywords as input and return < meaning, definition> pairs in a local repository. Fig. 4 shows a snapshot of results obtained from the Contextual Sense Extractor.

The contextual sense based rank of the web page and its back-links as computed on the basis of the contextual senses of the various keywords present in the web page is addressed as the 'Rank (CSense/WP)'. The web pages having the higher Rank (CSense/WP) value are considered to be more contextually related to that sense for a given query. The Similarly Rank (CSense/WP) of back-links to these web pages is computed to find out the more relevant back-links to be displayed to the user. Finally, web pages and back-links with significant rank values are displayed to the user and thus provide more relevant information in the desired area.

EDIT	KEYWORD_	CONTEXTUAL_SENSE
	mouse	mouse: any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails
	mouse	mouse: a swollen bruise caused by a blow to the eye
	mouse	mouse: person who is quiet or timid
	mouse	computer mouse: a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad
	mouse	pussyfoot: to go stealthily or furtively
	mouse	mouse: manipulate the mouse of a computer
	student	educatee: a learner who is enrolled in an educational institution
	student	student: a learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines
	business	business organisation: a commercial or industrial enterprise and the people who constitute it
		business: the activity of providing goods and services involving financial and commercial and industrial

Fig. 4. Contextual Senses from the WordNet Dictionary

The rank of a web page and its back-links (extracted by the back-link extractor) corresponding to different contextual senses of keywords can be computed based on the occurrence of various keywords in various HTML tags (a variation from the standard TD matrix of the IR, which computes only the total occurrence of a term in the document) and by assigning weights to different HTML tags. Using the steps listed below can accomplish this.

Step 1. Extract the keywords present in various HTML tags and assign weights as per their occurrences by considering the stemming of keywords, the removal of stop words, and noises as applicable.

Step 2. For each keyword 'ki' compute the accumulated weight using following expres sion.

$$AW_{k_i} = \sum_{j=1}^{n} N_j * W_j \tag{2}$$

Where,

AW_{ki}: the accumulated weight of keyword ki in the web page

N_j : the number of occurrences of keyword 'k_i' in tag 'j'

W_j : the weight of tag 'j'

n : the number of tags

The *keywords extractor* module performs Step 1 & Step 2.

Step 3. For the top 5 keywords, the *contextual sense extractor* is referred to for their contextual sense definition.

Step 4. Compute the conditional probability that a web page relates to a contextual sense (CSense) using the Bayesian rule [18, 19], which is given below:

$$P(CSense/WP) = \frac{P(CSense \cap WP)}{P(WP)} \tag{3}$$

Step 5. Assuming there is a set 'P' consisting of keywords that have been extracted from a web page (WP) that is represented as $k_{WP1}, k_{WP2}, \ldots\ldots k_{WPk}$ and another set 'Q' that consists of keywords that have been extracted from the contextual sense definition (taken from WordNet) that is represented as $k_{Def1}, k_{Def2}, \ldots.k_{Defn}$.; the probability is computed as:

$$P(CSense/WP) \propto \frac{P((k_{Def1,} k_{Def2,} \ldots\ldots\ldots k_{Defn}) \cap (k_{WP1,} k_{WP2,} \ldots\ldots\ldots k_{WPk}))}{\sum_{k=1}^{n} P(k_{WPk})} \tag{4}$$

Where,

WP: the Web Page

Def: the Definition

k_{WPi}: the i^{th} keyword in web page 'WP'
k_{Defi}: the i^{th} keyword in Definition 'Def'

Step 6. Let 'T' be the set of common keywords for a web page and the contextual sense definition. Given set 'T', the probability of occurrence of the contextual sense (CSense) in a web page can be conveniently used to compute the rank of the web page in respect to that CSense (by using the conditional probability in equation 3).

$$Rank(CSense/WP) = \frac{\sum_{k=1}^{q} AW_k \ni k \in T}{\sum_{k=1}^{n} AW_k} \quad (5)$$

Where,

$k \in T$

k: the keyword in web page 'WP'
T: set of common keywords in the web page and contextual definition under consideration
WP: the web page under consideration
CSense: the contextual sense definition under consideration
AW_k: the accumulated weight of the keyword 'k' present in web page 'WP'
q: the number of keywords in set 'T'
n : the number of keywords in the WP

Step 7. Repeat Step 4 for each contextual sense definition (obtained from WordNet) of the top 5 highest weighted keywords.

Step 8. The maximum of all the values computed in Step 4 is taken as the most suitable contextual sense of the web page.

From the algorithm, it is observed that the above ranking module computes the rank of a web page based on the probability measure for various contextual senses to be satisfied by the web page, rather than on the basis of the number of link-counts for that web page. In a similar way the rank for the back-links to a web page is also computed.

4. EMPIRICAL EVALUATION

The proposed methods of rank computation were tested on two datasets. In this section, the evaluation process and the results obtained thereafter are discussed.

In order to carry out experiments on ranking web documents, we first need a repository of large number of pages that are relevant to different topics. For this purpose, we crawled the top few thousands of URLs on each of the various topics such as, 'mouse', 'rodent', 'crawler', 'arachnid', 'network host', 'jaguar', 'business', 'spider', 'server', 'colt', 'java', 'bank', etc. from Google. These URLs are considered to be dataset A for testing the proposed ranking mechanism. The back-links for these URLs have been extracted using a back-link extractor that was imple-

mented by us in [11] and are considered to be dataset B.

We hypothesize that the contextual sense based ranking of web pages and back-link pages yields better results than conventional ranking. To test our hypothesis, the two sets of experiments were conducted, as shown below.

Objective of Experiment Set 1:

- To generate the list of URLs and back-links of URLs related to each topic.
- To rank the list of URLs and back-links generated in the previous step according to the proposed ranking mechanism.

Objective of Experiment Set 2:

- To use the list of URLs and back-links generated in Experiment Set 1 to compare the ranking results of the proposed method with the Page Rank ranking mechanism
- To evaluate the effectiveness of the proposed ranking mechanism by using the standard metric that is called 'Precision'.

4.1 Rank Computation of a Web Page Based on Contextual Senses

This section describes how the web documents are evaluated on the basis of the various contextual senses of the keywords. Due to the space constraints, the following URL, which was randomly selected, has been chosen to display the detailed results:

http://en.wikipedia.org/wiki/Mouse_(computing)

It has been observed from the local repository that a total of '2,273' different keywords (where the keywords' length is greater than '3') were found in the above web page. The top 5 highest weighted keywords found are 'mouse', 'mice', 'button', 'computer', and 'user'. The WordNet dictionary [15-17] database was referred to get the contextual senses. It has been observed that for the top '5' keywords, the various contextual senses found were '6', '6', '9', '2', and '3', respectively. (i.e., a total of '26' different contextual senses.)

For instance, the contextual senses and corresponding definition for the highest weighted keyword, 'mouse', are listed in Table 2.

Next, the web page is evaluated against '26' contextual senses. The rank (CSense/WP) that is particular to each sense was then computed as described in Section 3.1. The highest ranked sense of each keyword is then selected from the list of all of the contextual senses. Table 3 shows the rank for '6' different contextual senses of the keyword 'mouse'.

Table 2. Contextual senses and definition of the keyword 'mouse'

Sense	Definition
Mouse	Any kind of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.
Mouse	A swollen bruise caused by a blow to the eye.
Computer Mouse	A hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad. On the bottom of the device is a ball that rolls on the surface of the pad.
Mouse	A person who is quiet or timid.
Pussyfoot	To go stealthily or furtively.
Mouse	To manipulate the mouse of a computer.

Table 3. Contextual senses and the computed rank of the keyword 'mouse'

Contextual sense	Rank (CSense/WP)
Mouse: rodent	0
Mouse: swollen bruise eyes	0
Mouse: computer mouse	0.4312
Mouse: person	0
Mouse: pussyfoot	0
Mouse: to manipulate a computer mouse	0.283751

Table 4. Contextual Senses and Rank (CSense/WP) of top '5' keywords

Keyword	Contextual Sense	Rank(CSense/WP)
Mouse	Computer Mouse	0.5947
Mice	Computer Mouse	0.5947
Computer	Computing Device	0.1614
Button	Push button	0.0772
User	Person	0.0447

Table 3 shows that the highest computed rank is for the sense 'Computer Mouse'. On a similar basis, Table 4 listed the highest rank sense for the top '5' keywords.

It shows that the document has the highest probability measure (rank) '0.59' for the sense 'computer mouse', '0.16' for the sense computer as 'a computing device', and '0.08' for 'button' as in a 'push button' sense, and '0.04' for the 'user' as a 'person'. This shows that the document in question is related to the word 'mouse' in the context of a computer device.

On a similar basis, the rank of other documents related to the same topic, taken as input was computed with respect to the various contextual senses of the top '5' highest weighted keywords present in the respective document. The highest rank (CSense/WP) value obtained is chosen as the computed rank for each document. For instance, the top '10' URLs that were obtained to display to the user on the topic 'mouse' and their computed highest rank are listed in Table 5.

4.2 Rank Computation of Back-Links

As stated earlier, the proposed mechanism considers back-links to be an important source of information. So, after downloading the whole set of URLs and extracting their links' information, back-links to these URLs are extracted using the back-link extractor module (Section 3). These back-link URLs are considered as being another set of URLs for serving a user query in order to improve the results. As an experiment, we computed the contextual sense based rank for all of the extracted back-links of web pages. The highest value is considered to be a computed rank. A list of some randomly chosen extracted back-links (for links obtained in Table 5) and their computed rank are given in Table 6.

The results thus obtained are ordered in decreasing fashion of their computed rank. The top '10' links having a higher ranking value are selected as being the more relevant results that are to be displayed to the user and they are listed in Table 7. The 'order' field in the table represents the position of the link in the top '10' results. The field 'Link Type' indicates the type of link. (i.e. a web page or a back-link of some web page that is under consideration.)

Table 5. URLs and their computed rank using our proposed mechanism on topic 'mouse'

S.No.	URL	Computed Rank
1	http://en.wikipedia.org/wiki/Mouse_(computing)	.63
2	http://www.pranavmistry.com/projects/mouseless/	.468
3	http://simple.wikipedia.org/wiki/Computer_mouse	.459
4	http://www.microsoft.com/hardware/en-us/p/touch-mouse	.42
5	http://www.mouseprogram.com/	.301
6	http://www.howstuffworks.com/mouse.htm	.28
7	http://searchexchange.techtarget.com/definition/mouse	.20
8	http://www.apple.com/magicmouse/	.119
9	http://www.webopedia.com/TERM/M/mouse.html	.117
10	http://www.logitech.com/en-us/mice-pointers/mice/performance-mouse-mx	.059

Table 6. Back-links and their computed rank using our proposed mechanism

S.No.	URL	Back link (URL)	Computed Rank
1	http://www.webopedia.com/TERM/M/mouse.html	http://www.webopedia.com/TERM/I/integrated_peripheral.html	.23
2	http://www.apple.com/magicmouse/	http://www.456bereastreet.com/archive/200508/mighty_mouse/	.21
3	http://www.logitech.com/en-us/mice-pointers/mice/performance-mouse-mx	http://lifehacker.com/5885687/what-we-use-whitson-gordons-favorite-gear-and-productivity-tips	.11
4	http://en.wikipedia.org/wiki/Mouse_(computing)	http://www.dansdata.com/diamondback.htm	.43
		http://blog.futurestreetconsulting.com/2008/12/09/crowdsource-yourself/	.22
		http://www.tabletouch.com/monitor_details.html	.16
		http://en.wikipedia.org/wiki/Computer_case	.19
5	http://www.howstuffworks.com/mouse.htm	http://www.edinformatics.com/inventions_inventors/computer_mouse.htm	.068
6	http://www.mouseprogram.com/	http://www.lawrencegoetz.com/programs/java.htm	.12
		http://tech.worlded.org/docs/cesol/tutorials.htm	.4
		http://www.narlib.org/computer_class_info	.2
		http://publiclibrary.cc/computerinternettutorial.htm	.30
		http://www.lazeetek.com/html/links.html	.36
7	http://www.pranavmistry.com/projects/mouseless/	http://www.techvert.com/mouseless-invisible-computer-mouse/	.15
		http://designtaxi.com/news/32363/MIT-Develops-Invisible-Computer-Mouse/	.13
		http://imulus.com/blog/category/concepts/	.10
		http://www.bbc.co.uk/programmes/p008tz1c	.11
8	http://www.microsoft.com/hardware/en-us/p/touch-mouse	http://www.microsoft.com/hardware/en-us/bluetrack-technology	.31
		http://www.guomii.com/posts/5279	.28
		http://www.microsofthardwareblog.com/page/3/	.36
9	http://simple.wikipedia.org/wiki/Computer_mouse	http://en.wikipedia.org/wiki/Mouse_(computing)	.63
10	http://searchexchange.techtarget.com/definition/mouse	http://ajaxian.com/archives/mouse-entering-and-leaving-versus-over-and-out	.13
		www.whatis.com/mouse.htm	..2
		http://www.bitpipe.com/tlist/Mice.html	.22

Table 7. The top '10' high ranking URLs consisting of URL's and back-links

Link Type	URL	Computed Rank (CSense/WP)	Order (Using the CSB rank) (Position in the top '10' results)
Web Page	http://en.wikipedia.org/wiki/Mouse_(computing)	.63	1
Web page	http://www.pranavmistry.com/projects/mouseless/	.468	2
Web Page	http://simple.wikipedia.org/wiki/Computer_mouse	.459	3
Back-Link	http://www.dansdata.com/diamondback.htm	.43	4
Web Page	http://www.microsoft.com/hardware/en-us/p/touch-mouse	.42	5
Back-Link	http://tech.worlded.org/docs/cesol/tutorials.htm	.4	6
Back-Link	http://www.lazeetek.com/html/links.html	.36	7
Back-Link	http://www.microsofthardwareblog.com/page/3/	.36	8
Back-Link	http://www.microsoft.com/hardware/en-us/bluetrack-technology	.31	9
Web Page	http://www.mouseprogram.com/	.301	10

The results in Table 5 show that some of the URLs with a low contextual rank are also displayed, which may not be more relevant to a query keyword. Whereas, the results in Table 7 show that when both URLs and their back-links are considered for ranking, the lower contextual ranked URLs from Table 5 are replaced with higher contextual ranked back-links URLs. This occurs as a result user gets more relevant links earlier on the first few pages. For instance, a new higher ranked back-link page (http://www.dansdata.com/diamondback.htm) is inserted in the 4th position, which results in shifting the low ranked page "http://www.microsoft.com/hardware/en-us/p/touch-mouse" to the 5th position. Similarly, new high ranked back-links' pages are inserted at positions from 6 to 9 in Table 7. It is also important to note that our proposed mechanism only picks higher contextual rank back-link URLs from Table 6 for replacement.

4.3 Comparison of the Proposed Ranking Mechanism with the Page Rank Algorithm

The Page Rank algorithm computes the rank of a link (relevance rank) on the basis of the popularity of the pages that point to it. The pages on web sites that are popular and pointed to by other popular sites are ranked higher by the page rank algorithm than as compared to some new sites that are not that popular right now. Whereas, the proposed mechanism computes the rank depending upon the contextual senses of the keywords that are present within the page (i.e., the content of the web is considered to be a parameter). In order to compare the ranking of both of the approaches, Experiment Set 2 was conducted to compare the ordering of links done by our contextual sense based ranking mechanism with the ranking being done by a standard page rank algorithm [6, 9]. Both algorithms gives different rank values, so we are not comparing algorithms based on rank values. Rather, we are using rank order as a comparison parameter. Rank order is more significant than rank value to a user who generally looks for the top results that are more related to the query [20].

In Experiment Set 2 the results obtained in Table 7 in Experiment 1 are taken as the input for computing the page rank score. For the purpose of computing the page rank of the URLs the PR checker tool [27] has been used in this work. The page rank score thus obtained are sorted in a

Table 8. Page rank ordering

Link Type	URL	Computed Rank (CSense/WP)	Order (using the page rank score)
Web Page	http://en.wikipedia.org/wiki/Mouse_(computing)	0.63	1
Web Page	http://simple.wikipedia.org/wiki/Computer_mouse	0.459	2
Web Page	http://www.microsoft.com/hardware/en-us/p/touch-mouse	0.42	3
Back-Link	http://www.microsoft.com/hardware/en-us/bluetrack-technology	0.31	4
Web page	http://www.pranavmistry.com/projects/mouseless/	0.468	5
Web Page	http://www.mouseprogram.com/	0.301	6
Back-Link	http://tech.worlded.org/docs/cesol/tutorials.htm	0.4	7
Back-Link	http://www.lazeetek.com/html/links.html	0.36	8
Back-Link	http://www.microsofthardwareblog.com/page/3/	0.36	9
Back-Link	http://www.dansdata.com/diamondback.htm	0.43	10

decreasing order of rank values and are shown in Table 8. The 'Order' field in Table 8 specifies the position of the URL as decided by the page rank algorithm.

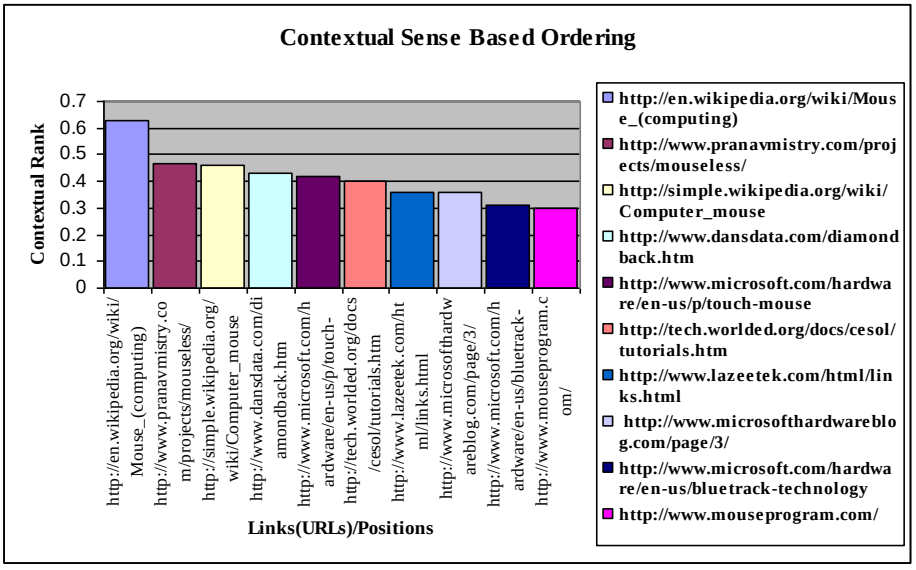
The results in Table 7 and 8 show that both of the algorithms have given different orders to the same set of URLs and back-link URLs.

Fig. 5(a) and Fig. 5(b) show the ordering of web pages done by two algorithms, the Contextual Sense Based (CSB) ranking algorithm and the Page Rank algorithm. The different ordering positions decided by two algorithms are shown using the color code for the links. Each similar color in both of the graphs in Fig. 5 (a & b) represents the same link. For instance, the color 'pink' represents the link 'http://www.mouseprogram.com/' in both graphs.

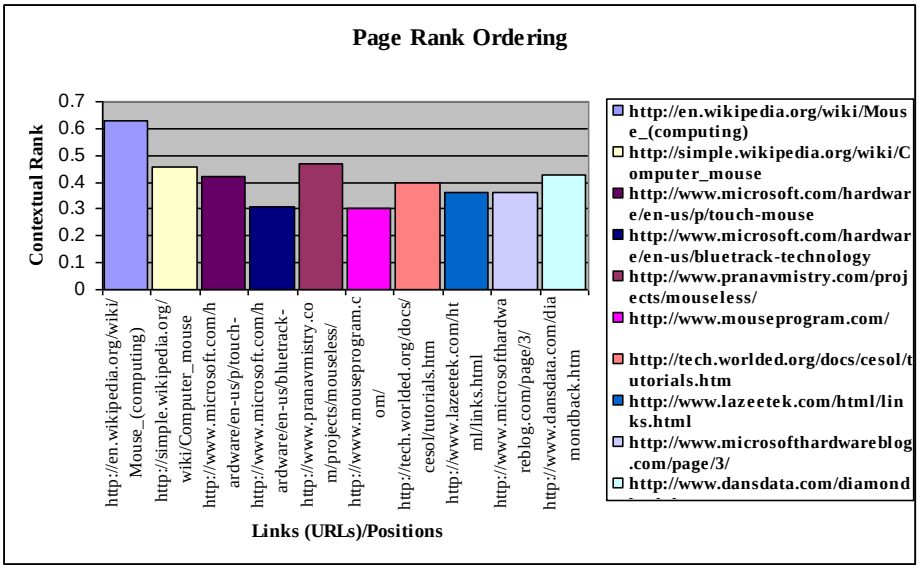
The ordering done by two algorithms is analyzed with the help of the change of color positions. The heights of bars represent the contextual relevance of the link. The following observations have been made:

a) Links with same contextual rank are placed in same order by both of the mechanisms In Fig. 5 (a & b) the same link has been placed at position '1' by both of the algorithms. The corresponding color bar has the same height in both graphs, which indicates the same contextual rank.
b) The links at positions 5, 7, 9, and 10 in the page rank ordering are displayed earlier by the CSB ordering at positions 2, 6, 8, and 4 respectively. This shows that contextually more related links are placed earlier by the CSB mechanism than by the page rank mechanism.
c) The page rank displays links 2,4 and 6 earlier, which all have a low contextual relevance. Whereas, CSB ordering placed these low relevance links at positions 3, 9, and 10 respectively. This shows that contextually low ranking links are placed later by CSB ordering.

These findings reflect that the CSB ranking algorithm displays contextually more related documents in the top positions and that it enables the user to get more related documents at higher positions than the page rank algorithm does. Hence, our method enables users to contextually see more related web pages in the first top positions.



(a)



(b)

Fig. 5. (a) Ordering Results by the Proposed Ranking Mechanism, (b) Ordering Results by the Page Ranking Mechanism

4.4 Evaluation Based on the Precision Metric

The evaluation of the ranking mechanism is done to ensure how accurately it ranks the web pages so that the more relevant web pages are ranked higher and displayed earlier. Generally,

the accuracy is measured in terms of two standard metrics called 'recall' and 'precision'. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved [23, 24]. The metrics can be described as follows:

$$\text{Re}\,call = \frac{relevnat\_webpages\_retrieved}{relevnat\_webpages\_retrieved + relevant\_webpages\_notretriev\,ed}$$

$$\text{Pr}\,ecision = \frac{relevant\_webpages\_retrieved}{relevant\_webpages\_retrieved + irrelevant\_webpages\_retrieved}$$

For the ranking of the web pages the recall cannot be taken as a standard metric to compare the results produced by two different ranking mechanisms, as the recall metric is dependent on the size of the dataset (i.e., the total number relevant web pages on a topic). The size of relevant web pages related to a topic on WWW is unknown. Therefore, in this work, the precision metric is chosen for performance analysis purpose. The precision values are computed for the top few results displayed to the user by the proposed ranking mechanism and by the page rank mechanism. The results thus obtained are as represented in the Fig. 6. Fig. 6 shows the average precision of the results from the proposed ranking mechanism compared to the page rank ranking mechanism. The proposed mechanism achieves higher precision overall. The result shows that in the top '10' results the proposed mechanism displays '8' related web pages, while by Google's page rank is only '6'.

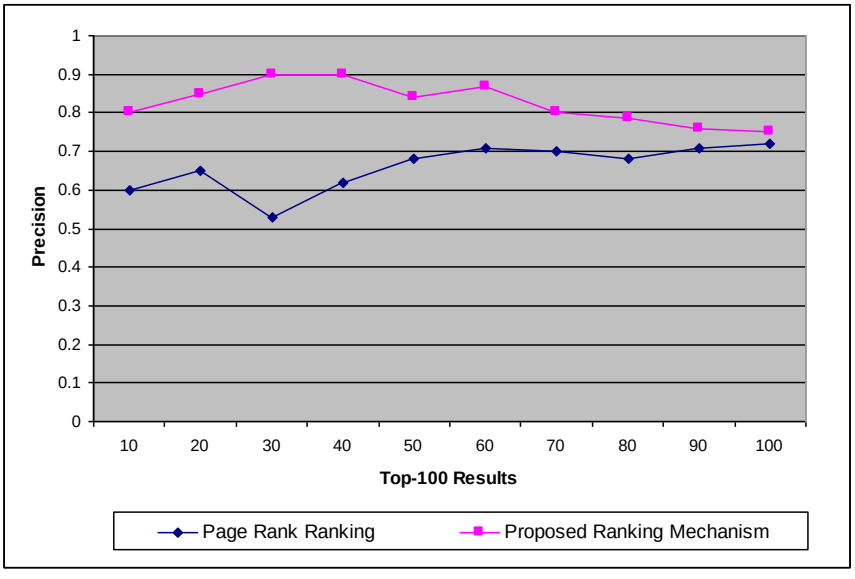


Fig. 6. Average Precision of the Results from the Proposed Ranking Mechanism Compared to the Page Rank Ranking Mechanism

Common Findings and Observations

We proposed a contextual sense based (CSB) ranking mechanism that uses the content and back-links to a web page as criteria to improve the ordering in which relevant documents to a desired user query can be found. The following two observations have been made from the experimental analysis:

- Introduction to back-links with web pages improves results with more contextually related web documents in response to a user query.
- The CSB ranking mechanism orders the more contextually related web pages into the top positions, as compared to the page rank algorithm. Thus, the user will find more related results earlier in the top positions.
- The CSB ranking mechanism results in higher precision as compared to Google's results. Thus, the user will get more related web pages earlier.

5. CONCLUSION

Existing ranking mechanisms mainly rank the web documents based on their link structure and popularity and do not consider the contextual senses of keywords for ranking. As a result, sometimes less relevant documents are displayed in the top positions. However, it has been observed that generally users mostly browse through the top 10 to 20 results in order to get the desired information [20]. So, there is a strong need to improve the ranking mechanism so that more relevant results are displayed earlier to the user.

For this, we have proposed a ranking mechanism that improves the ordering by using the contextual senses of keywords to compute the rank of a web document and its back-links. Back-links to matched web pages are extracted from the back-link extractor. The contextual senses of the web pages and their back-link pages are computed by the proposed mechanism. The web documents and their back-links are then ranked based on contextual senses and are displayed to the user. The experimental evaluation of the proposed ranking mechanism have shown that the overall search results in response to a user query are improved by replacing some of the contextually less relevant web pages with the contextually more relevant back link pages. Experiments have also shown that the proposed CSB ranking mechanism puts contextually more related web pages in the top order, as compared to the page rank algorithm, and this results in comparatively higher precision. Thus, the user will find more related results in the top positions earlier.

REFERENCES

[1] 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.

[2] 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.

[3] C. Ridings, M. Shishigin, "Page Rank Uncovered", Technical Report, September, 2002, http://www.voelspriet2.nl/PageRank.pdf

[4] P. Gupta, "Context based relevance evaluation of web documents", Proceedings of 5th International Conference, IC3 2012, Noida, India, August 6-8, 2012, pp.201-212.

[5] Kleinberg, J. M., "Authoritative sources in a hyperlinked environment", Journal of ACM, vol.46, no.5,

September, 1999, pp.604-632.
[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", Proceedings of 7th International WWW Conference, 1998, pp.107-117.
[7] Emil Gatail, Z. Balogh, "Focused Web crawling Mechanism based on Page Relevance", Proceedings of ITAT-Workshop on Theory and Practice of IT, Račkova dolina, Sept. 2005, pp.41-46.
[8] Z. Liu, Y. Du, Y. Zhao, "Focused Crawler based on Domain Ontology and FCA", Journal of Information & Computational science, Vol.8, no.10,2011, pp.1909-1917
[9] A. N. Langville, C. D. Meyer, "Deeper Inside Page Rank", Internet Math. J., Vol.1, No 3, 2005, pp.335-380.
[10] M. Persin, "Filtered document retrieval with frequency sorted indexes", Journal of the American Society for Information Science, Vol.47, No.10, October 1996, pp.749-764
[11] Pooja 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.
[12] G.Salton, "Developments in automatic text retrieval", science 253, 5023, 30 August, 1991, pp.974-979.
[13] M. Cutler, H. Deng, S. S. Maniccam, and W. Meng, "A New Study on Using HTML Structures to Improve Retrieval", Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence,1999, pp.406-409.
[14] Sun Kim and Byoung-Tak Zhang, "Genetic Mining of HTML Structures for Effective Web-Document Retrieval", Journal Applied Intelligence (ACM), Vol.18, No.3, May-June 2003, pp.243-256.
[15] Ingo Feinerer and Kurt Hornik, wordnet: WordNet Interface. R package version 0.1-8, 2011, http://CRAN.R-project.org/package=wordnet
[16] Mike Wallace, Jawbone Java WordNet API, 2007, https://sites.google.com/site/mfwallace/jawbone
[17] Christiane Fellbaum, WordNet: An Electronic Lexical Database. Bradford Books, 1998.
[18] Bayes's Rule, "Lecture 4: Conditional probability, Total Probability", www.stat.cmu.edu/--~cshalizi/36-220/lecture4.pdf
[19] Brandon Kountz, Ashwini Miryala, Kyle Scarlett, Zachary Zell, "Bayes Rule, Conditional Probability, independence", November, 2006. https://controls.engin.umich.edu/wiki/index.php/Bayes_Rule,_conditional_probability,_independence
[20] J. Bar-llan, M. Mat-Hossan and M. Levene, "Methods of comparing rankings of search engine results", The International Journal of Computer and Telecommunications Networking - Web dynamics, Vol.50, No.10, July 2006, pp.1448-1463.
[21] Chakrabarti S., Gibson, D. A., McCurley, K. S.(1999)," Surfing the Web Backwards", In the proceedings of 8th World Wide Web Conference.
[22] Gyngyi, Z., Garcia-Molina, H., Pedersen, J., "Combating web spam with trustrank", In: VLDB, pp.576-587. (2004)
[23] R. Jizba, "Measuring Search Effectiveness", 2007. http://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching_-_Recall_Precision.pdf
[24] Precision and recall-Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Precision_and_recall
[25] 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.
[26] Chattamvelli, Rajan, "Some generalizations of the PageRank metric", National conference on current trends in advanced computing, CTAC'10, Banglore, April 2010, pp.172-175.
[27] PR checker tool, www.prchecker.info/check_page_rank.php

Pooja Gupta

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 interestsare Search Engines, Crawlers and Focused Crawling

Dr. Sandeep K. Singh

Dr. Sandeep Kumar Singh is an Assistant Professor at JIIT in Noida, India. He has completed his Ph.D in (Computer Science and Engineering). He has around 11+ years' experience, which includes corporate training and teaching. His areas of interests are Software Engineering, Requirements Engineering, Software Testing, Web Application Testing, Databases, Internet and Web Technology, Object Oriented Modeling and Technology, Programming Languages, Distributed Computing, Model-based Testing, and Applications of Soft Computing in Software Testing. He is currently supervising 5 Ph.D's in Computer Science. He has around 15 Papers to his credit in different international journals and conferences.

Dr. Divakar Yadav

Dr. Divakar Yadav received his PhD. degree in Computer Science and Engineering from Jaypee Institute of Information Technology, Noida, India in Feb 2010. He spent one year, from Oct 2011 to Oct 2012, at Carlos III University, Leganes, Madrid, Spain as a post doctoral fellow. He has published more than 30 research papers in international/national journals and conferences. His areas of interest are Information retrieval and soft computing.

Prof. A. K. Sharma

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.