web to measure the quality of web pages.

Design of a Framework for Knowledge Based Web Page Ranking

PremSagar Sharma^{#1}, A. K. Sharma^{*2}, Pankaj Garg^{#3}

^{#1}Research Scholar, Uttarakhand Technical University Dehradun, U.K., India
 ^{*2} Department of Computer Science & Engineering, BSAITM Faridabad, H.R., India
 ^{#3}Computer Science & Engineering, Dev Bhomi Group of Institutions, Dehradun, U.K. India
 ^{#1}premsagar1987@rediffmail.com,*²ashokkale2@rediffmail.com, ^{#3}hodpankaj5@gmail.com

Abstract-Web is growing exponentially. The search mechanisms need to provide relevant and high quality web pages that too in short time to the internet user. The standard search engines utilize the link structure of the web to measure the quality of Web pages. Wherein it has been observed that the some less popular and lowly ranked but significantly important web pages remains missing. In this paper a framework for knowledge based web page ranking is being presented. It provides relevant and quality information in desirable time with the help of a proxy server. This framework exploits the content of the

Keywords — Introduction, web page ranking algorithms, proposed framework for knowledge base web page raking.

1. INTRODUCTION

Increasing popularity of internet has rendered World Wide Web, a rich collection of hypertext documents belonging to different domains. Web based information retrieval system called search engine, though has made things easy for information seeker but still it does not provide guarantee about the correctness of the information provided to the user. Many times the information is not precise.

Information retrieval systemprovides the information to the user based on certain retrieval criterion. For instance, it may search the web for identifying documents which contain information on a given subject. Due to the large size of the WWW, it is very common that a large number of documents get identified related to a particular domain. Therefore to help guide users towards finding the best matching documents, a ranking mechanism is employed by the search engine. Common methods for ranking are either based on relevancy where the documents are ordered from most relevant to least relevant or on the basis ofpopularity where documents are ordered from being most popular to least popular. It is important to understand that the term popularity is normally the result of link analysis and not user feedback. A web search engine typically consists of a ranking System that measure the importance of Web Pages (discussed in sec-1.2).

1.1 Web Mining

In web mining, the techniques of data mining are used to automatically discover and extract information from Web documents and web services. With a view to extract something useful out of the Web. The following tasks are generally performed for this purpose:

1) Resource finding: Useful resources are retrieved from the web documents i.e. we extract the data which are accessible on the web either through online or offline mode.

2) Information selection and pre-processing: Specific information is selected automatically and pre processing of that information is carried out with a view to data cleaning, normalization, feature extraction etc.

3) Transformation: The original retrieved data is transformed into information rejuvenation of stop words to obtain the necessary representation for finding phrases in training mass.

4) *Generalization*: The general patterns present in individual web sites or across multiple sites are found by generalization. Machine learning and data mining techniques are employed for this purpose.

5) Analysis: Validation and interpretation of the mined patterns is done in phase of analysis. It has got an important role for pattern mining. The human being plays an important role for knowledge discovery technique on the web.

1.2 Web Mining Taxonomy

Web mining is categorized into three different types as shown in Fig.-1[13].

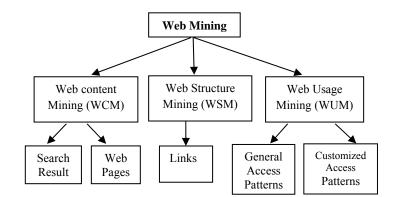


Fig. 1: Web Mining Taxonomy

A. Web Usage Mining (WUM)

The data obtained from the web server, access logs, browser logs, proxy server logs, user profiles, registration data, user queries, user sessions or transactions is called secondary data. The secondary data are mined through the web usage mining. Usage data grabs the identity or origin of Web users along with their browsing activities at a Web site. Web Usage Mining uses data mining techniques to realize important usage patterns from Web data in order to understand and to provide better services of Web-based applications.

B. Web Structure Mining (WSM)

Web structure mining discovers the structural synopsis about a web site and its underlying web pages. Through Web Structure Mining the link structure of hyper linkedresources are discovered. This model is used to classify and compareweb pages or integrate different web pages. Web structure mining is carried out in one of the following ways.

- 1. First extract the patterns from hyperlinks in the web: A hyperlink is a structural component that connects the web page to another web page at a different location.
- 2. Second mine the document structure: analysis of the tree-like structure of page structures to clarify HTML or XML tag usage.

C. Web Content Mining (WCM)

Web content mining extracts useful information such as text, images, audio, video, records from the contents of the web documents. Mining supplied web documents as well the result pages produced from a search engine. Basically there are two approaches in content mining called:

- 1. Agent based approach: The agent based approach relies on searching proper information using the uniqueness of a particular domain to interpret and organize the collected information.
- 2. Database based approach: The database approach is used to get back the semi structured data from the web.

1.3 Web page Ranking Algorithms

Web page ranking algorithms are categorized into two parts on the basis of web links and web content.

1- Content-based algorithms- these algorithms return all those web pages which are matching user query words with the web documents.

Vector space [2], TF-IDF [3] and BM25 [4] are examples of these algorithms. These algorithms are used for searching structured pages/documents within digital libraries, rather than the unstructured web pages.

- 2- *Connectivity-based algorithms* these algorithms work on the basis of link between web pages i.e. the importance and relevancy of web pages is computed on the basis of links. These algorithms are further categorized into two major parts:
 - ✓ *Query-independent-* PageRank [5], HostRank[6] and DistanceRank[7]. These algorithms use the entire web graph and compute the score of web pages offline.
 - ✓ *Query-dependent* HITS [8] Algorithm creates a query-specific graph online and thereafter computes as rank of the web pages.

2. RELATED WORKS

Wenpu Xing and Ali Ghorbani modified to original page rank algorithm and named it Weighted Page Rank (WPR) Algorithm [14], where in the rank score is decided based on the popularity of the pages. The rank of the pages is computed at the indexing time. It is providing high value of rank to the more popular pages. Every outlink page is given a rank value based on its popularity. Popularity of the page is decided on the basis of number of its in-links and out-links.

Weight Links Rank (WL Rank) algorithm [15] is the modification of standard page rank algorithm, given by Ricardo Baeza-Yates and sEmilio Davis. It assigns weight to web links based on three attributes: Relative position in page, tag where link is contained, length of anchor text. Relative position was not so effective, indicating that the logical position not always matches the Physical position.

HITS [16] is the oldest official Page Ranking algorithm which divides pages into two categories, *Authority*- the page which is pointed by many hyperlinks and *HUBs*- the page which points to various hyperlinks. It is primarily, a link based algorithm. Where in the web page is decided by analyzing their textual contents with respect to a given query string. Modified HITS (PHITS) is a modification of HITS where in a weight value is assigned to every link depending on the terms of queries and endpoints of the link [17]. A probabilistic explanation of relationship of term document is provided by PHITS.

TagRank (TR) Algorithm [18] is Web Content Mining algorithm for page ranking called TagRank (TR) algorithm. It is a comparison based approach along with on social comments which calculate the heat of the tags by using time factor of the new data source tag and the comments behavior of the web users.

In Time Rank algorithm (TIR) [19] the default rank of web page is computed on the basis of visiting time of the page and visiting time considered as a factor that shows the degree of importance to the users. This algorithm utilizes the time factor to increase the accuracy of the web page ranking.

EigenRumor (ER) Algorithm [20] is proposed for ranking the blogs. The rank scores of blog entries as decided by the page rank algorithm is often very low so it cannot allow blog entries to be provided by rank score according to their importance. To resolve these limitations, an EigenRumor algorithm is proposed for ranking the blogs.

Relation based algorithm [21] which is known as the most accurate page ranking algorithm among those that use Web Content Mining proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources.

Query Dependent Page Ranking (QDR) [22] is a powerful semantic search engine that takes into account keywords and return page onlyif both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.

In Distance Ranking Algorithm (DRA) [23] ranking is done base on the shortest logarithmic distance between two pages. It is intelligent ranking algorithm, proposed by Ali Mohammad ZarehBidoki and Nasser Yazdani. It is based on reinforcement learning algorithm. In this algorithm, the distance between pages is considered as a punishment factor. In this algorithm the ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them.

A critical look at the available literature reveals that the following issues need be addressed-

- 1. Need to identify the less popular and lowly ranked but important pages.
- 2. Perceived delay in response by the users for their request over the web
- 3. Need to minimize the problem of information overkill.
- 4. With the exponential growth of WWW and its coupled with the perceived delay by the users, it becomes imperative to prefetch the information sought by a particular group of users.

3. PROPOSED FRAMEWORK FOR KNOWLEDGE BASED WEB PAGE RANKING

In this work, a framework for knowledge based web page ranking as shown in Fig-2 is being proposed that computes the relevancy of a web page in response to a user query by exploiting a Proxy Server, which works between user and search engine, Proxy Server is a potential tool that can be suitably employed to intercept all requests to the search engine to see if it can fulfill the requests by itself. If not, then only it may forward the request to the search engine.

3.1 Components of Knowledge Based Web Page Ranking Using Web Mining

The proposed framework is composed of following components

- A. Query Processor
- B. Search Engine
- C. Proxy Server

- i. Mapper
- ii. Update Query Log
- iii. Web Page Analyzer
- iv. Knowledge Base
- A. *Query Processor*-Query Processor, analyzes the user query and converts it into aprecise query which understandable by the search engine which thereafter may return the most relevant set of results to the user. There are following steps followed by Query Processing:
 - 1- Tokenizing and Parsing
 - 2- Stop-word listing and QueryStemming
 - 3- Refine the user query to a search engine understandable form called precise query
 - 4- Pass the precise query to proxy server.

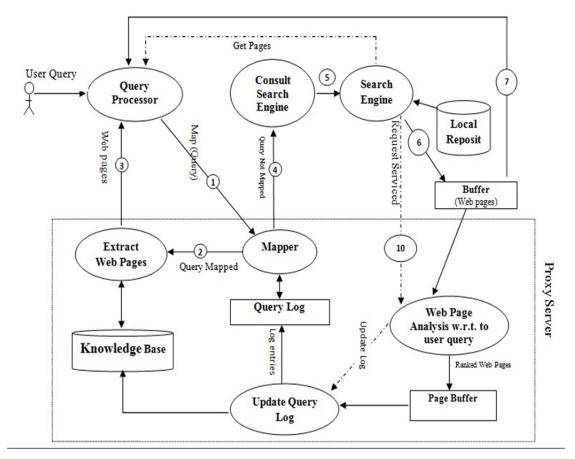


Fig.-2 - Framework for knowledge based web page ranking

- *B.* Search Engine- A search engine is the popular term for an information retrieval (IR) system. It allows the users to ask for content in the form of a query consisting of words/phrases and it retrieves a list of references that contain those keywords/phrases.
- *C.* **Proxy server** -Due to exponential increase of www, there are a large number of users that interact with servers through millions of networks connected with each other leading to a significant increase in the traffic on the internet. It is a potential tool that can be suitably employed to intercept all requests to the search engine to see if it can fulfill the requests by itself. If not, then only it may forward the request to the search engine. In fact, the proxy servers can be employed to achieve three main purposes:
 - 1. Increasing the relevancy of information
 - 2. Reduce perceived latency
 - 3. Minimizing the problem of information overkill.

Working Process of Proxy Server:

```
Algorithm: webpageAnalyzer (query)
{
        Step 1. Map 'query' in 'query log'
        Step 2.If(found)Extract corresponding webpages
        2.1 Return web pages.
        Step 3. If (not found)
                 3.1 Consult Search Engine for the query
                 3.2 Wait (request serviced)
                 3.3 Set web pages in to buffer
                 3.4 Perform web page analysis
                 3.5 Load ranked web pages in page buffer
                 3.6 Signal (update query log)
Algorithm: updatequeryLog()
while (1)
   {
      Wait (update query log)
      Set ranked web pages from page buffer
     Load the pages into knowledge base
     Load log entries into query log
   }
}
Mapper – mapper maps the query in the query log. If query is mapped then it invokes the web page extractor
```

otherwise consults with search engine for the web pages.

Query Log-The Query log is a data structure which stores successful queries fired over a period of time. Each query points to its resulted web pages stored in the database.

Query Log Updater - On receiving a signal from web page analyzer Query log updater updates the query log with the successful query.

Knowledge Base - Stores the ranked web pages pointed by query in query log.

Web page analysis – it is used to compute the relevance of a web page on the basis of its contents with respect to user query.

It follows following steps.

STEP-1: Extract web pages from buffer for a particular query and store them into Web Page Repository (WPR). The structure of WPR is given below.

Structure of web page repository (WPR)

	Title		Heading		Paragraph	
Doc_ID	Words	Freq_TK	Words	Freq_HK	Words	Freq_PK

Where:

Freq_TK – Frequency of a word in the title of a web page.

Freq HK - Frequency of a word in the heading of a web page.

Freq_PK - Frequency of a word in the paragraph of a web page.

STEP-2: Construct query keyword list (QKL) for a given query.

STEP-3:Compute the frequency of query keywords in title (Freq_T), heading (Freq_H) and paragraph (Freq_P) of each web page present in WPR and store the results in table of frequency of query keyword.

$$Freq_{Ti} = \sum_{j=0}^{m} Freq_{TKi(j)}$$

Where \mathbf{Freq}_{i} is frequency of all m query keywords in title of the ith web page in WPR.

$$Freq_Hi = \sum_{j=0}^{m} Freq_HKi(j)$$

Where $Freq_H_i$ is frequency of all m query keywords in heading of the ith web page in WPR.

$$Freq_Pi = \sum_{j=0}^{m} Freq_TKi(j)$$

Where **Freq_Pi** is frequency of all m query keywords in paragraph of the ith web page in WPR.

Structure of total Frequency of query keyword (FQK) is given below.

Structure	of Frequency	of FQK

Doc_Id rreq_I rreq_H rreq_P

STEP-4:Computes the total number of words in title (TW), heading (HW) and paragraph (PW) of a page present in WPR.

TWi = \sum Title_Words_i HWi = \sum Heading_Words_i PWi = \sum Paragraph_Words_i

Where:

 $Title_Words_i$, $Heading_Words_i$ and $Paragraph_Words_i$ are total frequency of word in title, heading and paragraph of the ith web page present in WPR respectively.

STEP-5: Compute the Total Relevancy (TRi) of ith web page

$$TRi = 1/m[(10*(Freq_T_i/TW_i)) + (2*(Freq_H_i/HW_i)) + (1*(Freq_P_i/PW_i)] - \dots (1)$$

STEP-6: Arrange the links of web pages in descending order based on Total Relevancy (TR_i)

STEP-7: Signal (request_serviced)

4. EXPERIMENTAL RESULTS

A user query "**human survival in society**" was fired on Google search-engine and 10 web pages obtained from thereof are listed in Table 1.

S.No.	Doc_ID	URL		
1	wp-1	www.personalityresearch.org/papers/smith.html		
2	wp-2	www.wsu.edu/~taflinge/socself.html		
3	wp-3	www.wsu.edu/~taflinge/biology.html		
4	wp-4	www.wsu.edu/~taflinge/culture1.html		
5	wp-5	https://www.quora.com/In-nature-only-the-fittest-survive-How-does-society-and- technology-affect-human-evolution-given-nearly-everyones-genes-are-getting- passed-on-to-future-generations-contrary-to-what-happens-in-the-wild		
6	Wp-6	https://www.quora.com/Does-survival-of-the-fittest-apply-to-humans-Are-the-best-specimens-of-our-species-really-outlasting-others		
7	wp-7	anthro.palomar.edu/culture/culture_2.htm		
8	wp-8	science.jrank.org/pages//Social-Darwinism-Human-Nature-Struggle-Survival.html		
9	wp-9	https://www.linkedin.com/pulse/20140702020006-232933375-basic-survival-tips-living-in-today-s-society		
10	wp-10	http://origins.osu.edu/article/climate-human-population-and-human-survival-what- deep-past-tells-us-about-future		

Table 1- Link of Web Pages

The proposed framework was tested on the 10 web pages (table -1) obtain from the Google search. The following keywords of the given query (**QKL**) were used to obtain the data about title, heading and paragraph words from the pages listed in table-1. The data obtained is provided in table-2 and table-3.

QKL Human survival society

Table -2: Total number of words in the Title, Heading and Paragraph of web pages present in WPR

<i>S.N</i> .	Doc _ID	Title _Words (TW)	Heading_Words (HW)	Paragraph _Words (PW)
1	wp-1	3	32	1176
2	wp-2	4	19	543
3	wp-3	4	13	992
4	wp-4	3	5	1190
5	wp-5	12	7	661
6	wp-6	4	0	1002
7	wp-7	23	27	227
8	wp-8	12	18	566
9	wp-9	9	9	611
10	wp-10	16	19	824

Table -3: Frequency of query keywords in the Title, Heading and Paragraph of web pages

Doc_Id	freq_T	freq_H	freq_P
wp-1	1	2	74
wp-2	1	1	40
wp-3	1	1	64
wp-4	1	1	70
wp-5	2	0	63
wp-6	1	0	28
wp-7	2	2	4
wp-8	3	1	22
wp-9	2	0	2
wp-10	3	0	42

The expression -1 was applied on data provide in table-2 and table -3 to obtain total relevancy and rank of each web page as listed in table -4.

Table 4- Total Relevancy of web pages & its rank

Doc_ID	Total Relevancy	Rank
wp1	1.173753	2
wp2	0.892976	4
wp3	0.906121	3
wp4	1.264052	1
wp5	0.587326	9
wp6	0.842648	6
wp7	0.345112	10
wp8	0.883327	5
wp9	0.741832	7
wp10	0.64199	8

It may noted that the page with high relevancy have been high rank where rank 1>2>3 - - . The same dataset of web pages was given to different experts to compute the rank of each web page manually and resulted ranks of the pages have been compared with the rank obtained from the proposed work as shown in table – 5.

Doc_ID	Search Engine	Manual	Rank
wp-1	1	1	2
wp-2	2	4	4
wp-3	3	3	3
wp-4	4	2	1
wp-5	5	8	9
Wp-6	6	6	6
wp-7	7	10	10
wp-8	8	5	5
wp-9	9	7	7
wp-10	10	9	8

Table 5- Comparison between system ranking and manual ranking of the web pages

Performance evaluation of the proposed mechanism is done based on Precision of the downloaded pages as given below.

Precision
$$p = \frac{r}{n}$$

Where: r- Number of relevant documents

n-Total number of documents

Comparison between the ranks obtained from a standard search engine (Google), manual rank from experts and rank provided by the proposed mechanism is given in table-5.

Precession for Google = 3/10 = 0.3

Precession for proposed mechanism = 6/10 = 0.6

Hence the performance of the proposed method is higher as compared to existing approaches.

5. CONCLUSION

The standard search engines return large number of web pages in response to user's queries, while the user always seeks relevant web pages that too in a short time. The page ranking mechanismplay import role in this direction as the search engine can choose the best ranked documents for the user.

In this paper we proposed a framework that computes page rank of a web page based on the knowledge based maintain at a proxy server. The performance of this mechanism was compared with the performance of standard search engine and manually obtains ranks from experts. And the comparison suggest that the result obtain thereof are better the performance of existing search engine.

REFERENCES

- NeelamDuhan, A. K. Sharma and Komal Kumar Bhatia (2009), PageRanking Algorithms: A Survey, proceedings of the IEEEInternational Advanced Computing Conference (IACC), pp 1530-1537.
- [2] G. Salton, C. Buckley (1998), Term-weighting approaches in automatic text retrieval, Journal of Information Processing and Management 24 pp. 513–523.
- [3] G. Salton (1971), ThesMART Retrieval System Experiments in Automatic Document Processing, Prentice-Hall, USA.
- [4] S.E. Robertson, S. Walker (1994), Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: Proceedings of SIGIR Conference on Research and Development in Information Retrieval, pp. 232–241.
- [5] L. Page, S. Brin, R. Motavni, T. Winogard (1998), The PageRank citation algorithm: bringing order to the web, Technical Report, Stanford Digital Library Technologies Project.
- [6] A.M. ZarehBidoki, N. Yazdani (2008), DistanceRank: an intelligent ranking algorithm for web pages, Journal of Information Processing and Management 44 pp. 877–892.
- [7] G.R. Xue, Q. Yang, H.J. Zeng, Y. Yu, Z. Chen (2005), Exploiting the hierarchical structure for link analysis, in: Proceedings of SIGIR Conference on Research and Development in Information Retrieval, pp. 186–193.
- [8] J.M. Kleinberg (1999), Authoritative sources in a hyperlinked environment, Journal of the ACM 46, 668-677.
- [9] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi (2012), Relevance Ranking and Evaluation of Search Results through Web Content Mining, In proceedings of the International Multi Conference of Engineers and Computer Scientists vol.1.
- [10] Pooja Sharma et al.(2010), Weighted Page Content Rank for Ordering Web Search Result, In proceedings of the International Journal of Engineering Science and Technology Vol. 2 (12), pp.7301-7310.
- [11] Ricardo Baeza-Yates and Emilio Davis (2000), Web page ranking using link attributes, In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329.

- [12] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S.Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg (1999), Mining the Web's Link Structure, Computer, 32(8), PP.60–67.
- [13] ThirumalaSreeGovada, N Lakshmi Prasanna (2014), Comparative study of various Page Ranking Algorithms in Web Content Mining, International Journal of Advanced Research Volume 2, Issue 7, pp. 457-464.
- [14] Wenpu Xing and Ali Ghorbani (2004), Weighted PageRank Algorithm, In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314.
- [15] Dilip Kumar Sharma, A. K. Sharma (2010), A Comparative Analysis of Web Page Ranking Algorithms, International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 08, 2010, 2670-2676.
- [16] Jon Kleinberg (1998), Authoritative Sources in a Hyperlinked Environment, In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.
- [17] Cohn, H. Changb (2000), Learning to Probabilistically Identify Authoritative Documents, In Proceedings of 17th International Conference on Machine Learning, PP. 167–174.Morgan Kaufmann, San Francisco, CA.
- [18] ShenJie, ChenChen, ZhangHui, SunRong-Shuang, Zhu Yan and He Kun (2008), TagRank: A New Rank Algorithm for Webpage Based on Social Web, In proceedings of the International Conference on Computer Science and Information Technology.
- [19] H Jiang et al. (2008), TIMERANK: A Method of Improving Ranking Scores by Visited Time, Seventh International Conference on Machine Learning and Cybernetics, Kunning, 12-15.
- [20] Ko Fujimura (2005), Takafumi Inoue, Masayuki Sugisaki, The EigenRumor Algorithm for Ranking Blogs, In WWW 2nd Annual Workshop on the Weblogging Ecosystem,
- [21] FabrizioLamberti, Andrea Sanna, Claudio Demartini (2009), A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines, In IEEE Transaction of KDE, Vol. 21, No. 1.
- [22] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee (2009), A Query-Dependent Ranking Approach for Search Engines, Second International Workshop on Computer Science and Engineering, Vol. 1, PP. 259-263.
- [23] Ali Mohammad ZarehBidoki (2007), Nasser Yazdani, DistanceRank: An Iintelligent Ranking Algorithm for Web Pages, Information Processing and Management.

AUTHOR PROFILE

PremSagar Sharma is B.Sc, MCA, M.Tech. (CSE) and pursuing Ph.D in Computer Science & Engineering fromUttrakhand Technical University Dehradun, India. Currently working as Assistant Professor, B. S. Anangpuria Institute of Technology and Management Faridabad, Haryana since 2015. He has published more than 10 papers international journals and international conferences. His area of research include Search Engine Optimization, Page ranking, Crawling, Indexing, Web mining etc.

Prof. A. K. Sharma received his M.Tech. (CST) with Hons.from University of Roorkee (Presently I.I.T. Roorkee) and Ph.D (Fuzzy Expert Systems) from JMI, New Delhi and he obtained his second Ph.D. in Information Technology form IIITM, Gwalior in 2004. Presently he is working as the Professor (CSE) & Dean (PG & Research) at BSAITM, Faridabad His research interest includes Fuzzy Systems, OOPS, Knowledge Representation and Internet Technologies. He has guided 28 Ph.D. thesis and 4 more are in progress with about 300+ research publications in International and National journals and conferences. The author of 8 books, is actively engaged in research related to Fuzzy logic, Knowledge based systems, MANETS, Design of crawlers. Besides being member of many BOS and Academic councils, he has been Visiting Professor at JMI, IIIT&M, and I.I.T. Roorkee.

Prof. (Dr.) Pankaj Garg is M.Tech. (CSE) and Ph.D. (Computer Science & Engineering). Currently working as a Professor atDev Bhomi Group of Institutions, Dehradun. His area of research include Wireless Communications, High-speed communication/computer networks, routing and scheduling, network security, Search Engine Optimization etc.