Volume 1, No. 5, December 2010
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

PRTF: PERSON RELATED TO A FIELD PROTOCOL FOR SEARCHING IN SOCIAL NETWORK DATABASES

Ashok Bhansali*[1] and Dr. H.R. Sharma*[2]

*[1] Associate Professor, Department of CSE, OP Jindal Institute of Technology, Raigarh, C.G., India
ashok.bhansali@opjit.edu.in

*[2]Director, Chhatrapati Shivaji Institute of Technology, Durg, C.G., India
hrsharma44@gmail.com

Abstract: Social networks are the contemporary ways to connect people across the globe. Social networking websites contain huge amount of data inside them. Volume of data is enormous and growing at a very fast rate. Social network data can be classified in three major categories – user profile data, user communication data and group communication data. Data mining can be applied effectively to discover the knowledge and to extract the useful patterns from this gigantic data set, which is called as the social network mining. In this paper we proposed a new search protocol to mine the information across all the social networking data in general, and use the extracted pattern to search an expert in particular. Further a mechanism to rank the searched experts is also proposed. Using this proposed protocol, apart from expert identification, number of useful patterns can be discovered from social networking data.

Keywords: Social network mining, data mining, expert finding, PRTF.

INTRODUCTION

Social Network

A social network is a social structure made of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige [Wikipedia].

Social networking is built on the idea that there is a determinable structure to how people know each other, whether directly or indirectly". Principle of "six degree of separation" is used to connect the people to each other without even knowing each other. Social networking often involves grouping specific individuals or organizations together. Large numbers of social networking websites exist in the www domain and newer ones are being launched continuously. Initially these websites were developed only for connecting the known persons and friends but today numbers of variants exist to serve the varied requirements of the mass. Special purpose social networking websites are also available to fulfill the specific requirements. While there are a number of social networking websites that focus on particular interests, there are others that do not.

Examples of Social Networking Based Applications

Flickr (www.flickr.com) - is a photo-sharing site based on a social network. Flickr exports an API for third-party developers, and we used this API to conduct the crawl. We also obtained group membership information via Flickr's API.

LiveJournal (www.livejournal.com) - is a popular blogging site whose users form a social network. LiveJournal offers an API that allows us to query for both forward and reverse links. We can also obtain group membership information via LiveJournal's API.

Orkut (www.orkut.com) – is a social networking site run by Google. Orkut is a "pure" social network, as the sole purpose of the site is social networking, and no content is being shared. In Orkut, links are undirected and link creation requires consent from the target. Because Orkut does not export an API, we had to resort to HTML screen-scraping to conduct our crawl, which requires more bandwidth. We obtained group information in a similar manner. Furthermore, Orkut limits the rate at which a single IP address can download the information and requires a logged-in account to browse the network.

YouTube (www.youtube.com) - is a popular video sharing site that includes a social network. Similar to Flickr, YouTube exports an API, and we used this feature to conduct our crawls. YouTube allows links to be queried only in the forward direction, similar to Flickr.

Facebook (www.facebook.com) - is a social networking website launched in February 2004 and operated and owned privately by Facebook, Inc. Users can add people as friends, can send them messages and update their personal profiles to notify friends about themselves.

LinkedIn (www.linkedin.com) - is a business oriented social networking site. Founded in December 2002 and launched in May 2003, it is mainly used for professional networking. As of 8 April 2010, LinkedIn had more than 65 million registered users, spanning more than 200 countries and territories worldwide.

Google Wave - Give It Time? Google Wave (http://wave.google.com/help/wave/about.html) is another form of social network. Google Wave is an online software application product from Google, which Google described

as "a new web application for real-time communication and collaboration".
There are many more social networking websites with many different purposes / intentions and newer ones are being developed continuously.

Social network mining

Social network mining refers to collect, extract and mine user profiles and various types of user generated data existing into the social networking sites. With the success and popularity of various social networking sites and huge data available into these sites, it's challenging to extract the useful patterns for various social, economic, education, business etc purposes**.**
Social network analysis software is used to identify, represent, analyze, visualize, or simulate nodes (e.g. agents, organizations, or knowledge) and edges (relationships) from various types of input data (relational and non-relational), including mathematical models of social networks. Network analysis tools allow researchers to investigate representations of networks of different sizes from small (e.g. families, project teams) to very large (e.g. the Internet, disease transmission). The output data can be saved in external files. Various input and output file formats exist. The various tools provide mathematical and statistical routines that can be applied to the network model.

Commonly used social networking protocols

FOAF (an acronym of Friend of a Friend) is a machine-readable ontology describing persons, their activities and their relations to other people and objects. Anyone can use FOAF to describe him or herself. FOAF allows groups of people to describe social networks without the need for a centralized database. It is a descriptive vocabulary expressed using the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Computers may use these FOAF profiles to find, for example, all people living in a particular region, or to list all people who you know and a friend of yours know. This is accomplished by defining relationships between people. Each profile has a unique identifier (such as the person's e-mail addresses, a Jabber ID, or a URI of the homepage or weblog of the person), which is used while defining these relationships. During recent years FOAF has been not only attracting more and more industry attention, but research interests as well.
XHTML Friends Network (XFN) is an HTML microformat developed by Global Multimedia Protocols Group that provides a simple way to represent human relationships using links. XFN was the first microformat, introduced in December 2003. XFN enables web authors to indicate relationships to the people in their blogrolls by adding one or more keywords as the *rel* attribute to their links.

RELATED & PREVIOUS WORK DONE

Data mining is applied on social networks by Jensen and Neville [1]. They have focused exclusively on the task of learning probability distributions over the values of attributes of objects and links. Classification of information sources for social network extraction is proposed by Kirchoff, Slabeva, Nicolai and Fleck [2]. The proposed classification scheme can be used for enhancing the information retrieval from a social network. The features of illicit group information with legitimate group data is compared and studied by Mukherjee and Holder [3]. They described how the graph-based knowledge discovery system, SUBDUE, when run in unsupervised discovery mode, finds structural patterns embedded within social network data. The feasibility of applying link-based methods in new applications domains is investigated by Agrawal, Rajagopalan, Srikant and Xu [4]. An algorithm named ComTector(Com-munity DeTector) is proposed by Du, Wu, Pei, Wang and Xu [5]. The algorithm is proposed to improve the efficiency for the community detection in large-scale social networks based on the nature of overlapping communities in the real world. A number of applications are proposed to which ComTector is applied. An analysis of the link structure of a general-purpose question answering community is proposed by Jurczyk and Agichtein [6] to discover authoritative users. Various issues affects web mining techniques are studied by Ting [7] for analysis of on-line social networks. Techniques and concepts of web mining and social networks analysis are introduced and reviewed along with a discussion about how to use web mining techniques for on-line social networks analysis. Social networks have the surprising property of being "searchable". A model is presented by Watts, Dodds and Newman [8] that offers an explanation of social network searchability in terms of recognizable personal identities i.e. sets of characteristics measured along a number of social dimensions.
The post-processing related work is studied in [9-13]. Organizing Web search results into clusters facilitate users' quick browsing through search results. Traditional clustering techniques are inadequate since they don't generate clusters with highly readable names. The clustering problem is mapped to a salient phrase ranking problem by Zeng, Chen1, Wei-Ying Ma, Qi-Cai He and Jinwen Ma [9]. The IR community has explored document clustering as an alternative method of organizing retrieval results, but clustering has yet to be deployed on most major search engines. An interface named Grouper is introduced by Zamir, and Etzioni [14] to the results of the HuskySearch meta-search engine, which dynamically groups the search results into clusters labeled by phrases extracted from the snippets. The proposed interface is also compared with the user Web search behavior on a standard ranked-list presentation versus a clustered presentation.
A typical newsgroup posting consists of one or more quoted lines from another posting followed by the opinion of the author. This social behavior gives rise to a network in which the vertices are individuals and the links represent "responded-to" relationships. An interesting characteristic of many newsgroups is that people more frequently respond to a message when they disagree than when they agree. This behavior is in sharp contrast to the www link graph, where linkage is an indicator of agreement or common interest. By analyzing the graph structure of the responses, it is possible to effectively classify people into opposite camps. In

contrast, methods based on statistical analysis of text yield low accuracy on such datasets because the vocabulary used by the two sides tends to be largely identical, and many newsgroup postings consist of relatively few words of text.

The expert finding problem is investigated in detail by Yimam [14] and the existing systems are reviewed and analyzed in this domain, and suggested a domain model that can serve as a basis for design and development decisions. An approach is proposed which is a method to generate, maintain and utilize an expertise information space based on dynamic organizational information resources. A methods for finding experts (and their contact details) using e-mail messages is proposed by Balog and Rijke [15]. The messages on a topic are located, and then the associated experts are discovered. Evaluation of the proposed method is done using the e-mail lists in the W3C corpus. The task of automatically determining an expert profile of a person from a heterogeneous corpus made up of a large organization's intranet is proposed by Balog and Rijke [16]. A profiling algorithms is also suggested which is applied to enhance the performance of a expert finding algorithm. The task of mining the email archives available in social network is done by Bird, Gourley, Devanbu, Gertz and Swaminathan [17]. The results from our data analysis over social networking email data are illustrated and newer patterns are discovered.

A conceptual framework is presented by Breslin, Bojar, Meza, Boley and Mochol [18] for the reuse and interlinking of existing, well-established vocabularies in the Semantic Web. The proposed framework can be used to connect people based on joint or complementing interests. Experts can be discovered using the profiles of people in social networks and using the content they post in online communities. The FindXpRT project for finding experts via rules and taxonomies is developed and proposed by Li, Boley, Bhavsar and Mei [19]. They have implemented rules for a client finding an expert to collaborate with, for an expert's decision making on whether to collaborate, and for specifying the collaboration mode. Collaborative environments are only effective when experts are accessible within them and those experts are able and willing to share their knowledge. Numbers of tools are described and studied by Maybury, D'Amore and House [20] which address some of the fundamental aspects of expertise management.

The issue of expert finding in a social network is addressed by Zhang, Tang and Li [21]. A propagation-based approach is proposed that takes into consideration both person local information and network information (e.g. relationships between persons). One important issue in sharing experience is to select relevant sharing partners who have appropriate knowledge and information on current specific topics. A method is proposed by Mori and Ishizuka [22] to employ the user profile and social structure of a web community in order to find sharing partners who have appropriate expertise and are likely to be able to reply to a request. The expertise investigation technique in online communities is proposed by Jiao, Yan, Zhao, and Fan [23]. The experts are discovered by using discussion groups and an expert ranking algorithm is also proposed with the work.

PROBLEM IDENTIGICATION

Formally, a social network can be defined as a graph $G = (V, E)$, where $v \in V$ represents a person in the social network and $e^t_{ij} \in E$ represents a relationship with type t between persons v_i and v_j. (t can be, for example, coauthor or colleague). The task of expert finding is defined as: given a query topic q, it is to *find* a subset of the persons from the social network and return them in a *ranked* list

The social networking data can broadly be classified in three major categories

A. *User Profile Data* - It includes all the data entered by user for the information of the user. Typically it is the personal information of the user - like first name, last name, address, gender, qualification, chat-ids, email-ids etc.
B. *User Communication Data* - The communication data includes number of things. The most important are the list of all the directly known and connected person's information and communication data with these connected person. The communication can also be done with the non-connected persons in a few social networking sites.
C. *Group Communication Data* - This is the data which belong to a particular group in the social networking site. All the members of the group can do the communication on some of the topics. This is called as the group communication data.

FOAF and other protocols have standard formats and mechanism to store and manipulate the social networking data. The only problem with these protocols is that there is no way to search the data in all the parts of social networking database and integrate the search results as per the requirements. Also there is no ranking system for the search results on the basis of which the results can be ordered properly and displayed to the user in some proper format. Hence there is a need of system where the searching can be done in user profile data, user communication data and group communication data and ranking of the search results for ordered display of the search results.

As apart from profile data, very useful and important information and trends could be hidden in to the user communication data and group communication data and so in the proposed protocol we include all the above three types of data for mining the Social Networks..

PROPOSED PRTF PROTOCOL

PRTF (Person Relation to a Field) is a protocol used to mine the people's information available across the social networking databases. A person participating in social networking can be a part of many different segments of the social networks, e.g. user can maintain his profile, can participate in a number of associated communication forums and can be a member of specific groups. The proposed protocol integrates different parts/data of the social networks to expand the search operation across all the segments of the social networking databases i.e. user profile, user communication and group databases. It is a search protocol

to generate the result as a set of people involved in a particular area or field available across social networking databases. The meaning of a "Field" in the protocol is any search key, which is used to search the persons (or people). The protocol also proposes ranking function for the searched results to produce the ranked list of the people found.

PRTF protocol can be used (integrated) with the social network mining to make social mining more useful. PRTF protocol gives a new dimension to the social network mining. Following are some of the examples where the PRTF can support social network mining:

A. The PRTF protocol can be used for finding experts. If the field (search) keyword is from any technical area then it can be seen as the technique to find an expert in the particular technical area. This method of expert finding is different from other methods, since the searching of experts is done across all the segments of the social networking databases. Finally the search result data is integrated and ranked as per the requirement.
B. The PRTF protocol can also be used for creating new groups and new forums on the basis of some search fields. For example suppose some one is searching for the field "Movie", then the set of persons retrieved as the outcome of PRTF protocol can be used to form new group called "Movie Group" or a forum can be created like "Movie Forum" where the persons retrieved can act as participant members and person with top rank can act as the moderator / administrators etc.

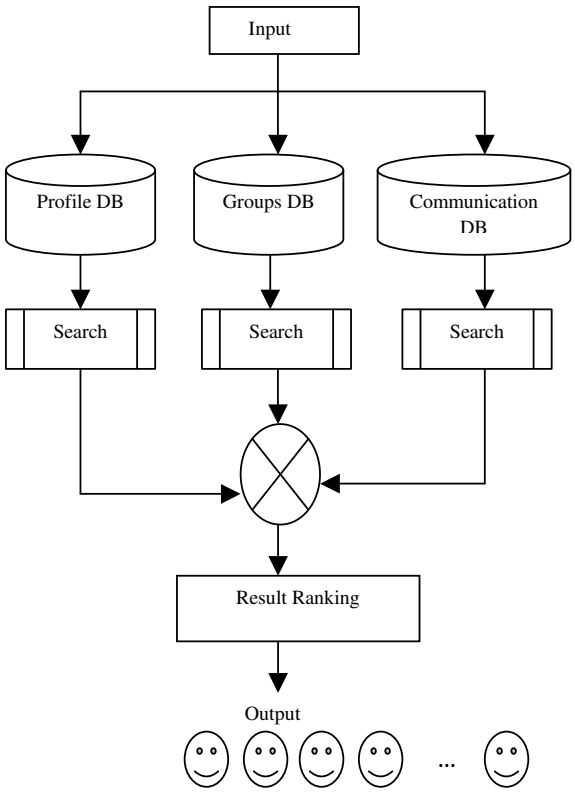

Figure 1. The PRTF Protocol Working Mechanism.

Fig-1 depicts the overall working mechanism of PRTF (Person Related To a Field) protocol. In the proposed work three major tasks are focused. First, searching across all the segments of social networking database. Second result ranking of the searching results, and lastly output representation of the ordered result is included in the proposed task.

The pseudo-code of the PRTF algorithm is given below, which is a five step algorithm. In the first step searching for a particular field is done in the profile data segment of the social networking sites. Then in the second and third step searching is done in the Group and Communication (Forum, Discussion etc.) databases of the social networking sites. In the fourth step search result for the first three steps is combined using the common entities available for the search, and finally the result rank is generated at step five.

```
Algorithm Algorithm for finding a person
related to a field (PRTF)
Require: Profile Data (PDB), Group Data
(GDB), Communication Data (CDB), Search
filed f
1. Extract features from Profile
Database
for all users u in PDB do
      - fetch the list of profile
      entities
      - Get the values of related
      entities ∈ f
      for all entities pe in the profile
      PDB do
      u.score(PDB) += u.score(pe)
      end for
end for

2. Extract features from Group Database
for all users u in GDB do
      - fetch the list of profile
      entities
      - Get the values of related
      entities ∈ f
      for all entities ge in the profile
      PDB do
      u.score(GDB) += u.score(ge)
      end for
end for

3. Extract features from Communication
Database
for all users u in CDB do
      - fetch the list of profile
      entities
      - Get the values of related
      entities ∈ f
      for all entities ce in the profile
      PDB do
      u.score(CDB) += u.score(ce)
      end for
end for
```

```
4. Combine the score of a user from all
the databases.
for all users u do
    u.score = u.score(PDB)+ u.score(PGB)
                          + u.score(PCB)
end for
5. Generate Ranks of the user using the
scores of the users
return list of top ranked users
```

ILLUSTRATION OF PRTF PROTOCOL

Suppose we are searching for a java expert in the social networking site. So the searching can be done into the profile information of the users, the group data of the user and communication data of the user into the various threads/blogs etc where s/he is a member.
The important task here is the ranking of searched user in the social networking site. Here in this example we want to rank the persons who are related with the field java. Information related to the java can be discovered at the user's profile information, user communication information whenever s/he is discussing about java with her/his friends and user's group data in the java related communities, if the user has joined these.

Once searching is done and persons are identified, the next step is to rank these persons in some order. This ranking can be done on the basis of ranking of the user's profile data, communication data and group communication data. Typically the rank function is derived from these three database searches

Result-Rank = f (profile-data, communication-data, group-communication-data)

Proper weights can be assigned for each database search and these weighted ranks can then be clubbed together to rank a particular search in all the three databases.

This Result-Rank can be explained using the following examples. Suppose, A and B are two persons and following are the data available for them at a social networking site:

		Person-A	Person-B
Profile Data	***Exp. In Java***	10 Years	5 Years
	Work Area	Java, Swings	Java, JSP, Servlet
Communication Data	***Java Threads***	30	60
Group Communication Data	***# of Java Communities Membership***	5	3

If we only look at the profile data of the user then A is more related to java then B. Hence in search ordering A should appear before B. But if we consider all the three databases namely profile data, communication data and groups communication data then things are no more same. A is a member of 5 java groups and B is a member of 3 java groups but the contribution of A's in the java related threads / blogs etc is 30, whereas there are 50 communication threads where B has participated.

If we assign higher weightage to the group search data then B is ranked better, related to the field java, than A. Hence in search ordering B should appear before A, which is opposite to the previous case where only profile data was considered.

CONCLUSIONS AND FUTURE WORK

In this paper we introduced the concept and framework for a new PRTF (Person related to a field) protocol. Protocol searches through the different types of databases available in the social networking sites and extracts the useful patterns. The main usage of the proposed protocol is exhaustive search in the social networking databases and ranking the results of a search. Apart from finding the expert the proposed protocol can be used for many different purposes like finding new communities, new patterns new groups etc. The future scope of the work could be adding generic implementation of extraction of social networking databases and designing the framework for the protocol including the visualization of the searched results.

REFERENCES

[1] David Jensen and Jennifer Neville , "Data Mining in Social Networks", In National Academy of Sciences Symposium on Dynamic Social Network Modeling and Analysis, 2002.
[2] Lars Kirchhoff, Katarina Stanoevska-Slabeva, Thomas Nicolai & Matthes Fleck, "Using social network analysis to enhance information retrieval systems", Applications of Social Network Analysis (ASNA) (Zurich), 2008.
[3] Maitrayee Mukherjee, Lawrence B. Holder, "Graph-based Data Mining on Social Networks", The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22-25 August 2004.
[4] Rakesh Agrawal, Sridhar Rajagopalan, Ramakrishnan Srikant, Yirong Xu, "Mining Newsgroups Using Networks Arising From Social Behavior", International World Wide Web Conference archive, Proceedings of the 12th international conference on World Wide Web table of contents, Pages: 529 - 535, 2003.
[5] Nan Du, Bin Wu, Xin Pei, Bai Wang and Liutong Xu, "Community Detection in Large-Scale Social Networks", International Conference on Knowledge Discovery and Data Mining archive, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on

Web mining and social network analysis table of contents, San Jose, California , Pages: 16-25, 2007.

[6] Pawel Jurczyk and Eugene Agichtein, "Discovering Authorities in Question Answer Communities by Using Link Analysis", Conference on Information and Knowledge Management archive, Proceedings of the sixteenth ACM conference on Conference on information and knowledge management table of contents, Pages: 919-922, 2007.

[7] I-Hsien Ting , "Web Mining Techniques for On-line Social Networks Analysis", Service Systems and Service Management, 2008, Page(s): 1 - 5, 2008.

[8] Duncan J. Watts, Peter Sheridan Dodds, M. E. J. Newman, "Identity and Search in Social Networks", Vol. 296. no. 5571, pp. 1302 – 1305, Science 17 May 2002.

[9] Hua-Jun Zeng, Zheng Chen1, Wei-Ying Ma, Qi-Cai He,Jinwen Ma, "Learning to Cluster Web Search Results", SIGIR'04, July 25–29, 2004.

[10] Rabia Nuray-Turan, Zhaoqi Chen, Dmitri V. Kalashnikov, Sharad Mehrotra, "Exploiting Web querying for Web People Search in WePS2", In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, April, 2009.

[11] Oren Zamir and Oren Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results", pp. 283-296, Eighth International World Wide Web Conference, Elsevier, Toronto, Canada, ISBN 0444502645, May 1999.

[12] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson and Samuel Ieong, "Diversifying Search Results", WSDM '09 Barcelona, Spain, 2009.

[13] FAST Search Best Practices, www.fastsearch.com, "Search query processing", 2006.

[14] Dawit Yimam, "Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach", ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop, 2000.

[15] Krisztian Balog Maarten de Rijke, "Finding Experts and their Details in Email Corpora", In International World Wide Web Conference, Proceedings of the 15th international conference on World Wide Web. Pages, 1035-1036, ISBN 1595933239, 2006

[16] Krisztian Balog and Maarten de Rijke, "Determining Expert Profiles (With an Application to Expert Finding)", International Joint Conference On Artificial Intelligence archive, Proceedings of the 20th international joint conference on Artifical intelligence, Pages: 2657-2662, 2007.

[17] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, Anand Swaminathan, "Mining Email Social Networks", MSR'06, May 22–23, 2006, Shanghai, China.

[18] John G. Breslin, Uldis Bojars, Boanerges Aleman-Meza, Harold Boley, Malgorzata Mochol, Lyndon JB Nixon, Axel Polleres, and Anna V. Zhdanova, "Finding experts using Internet-based discussions in online communities and associated social networks", In Proceedings of the 1st International ExpertFinder Workshop Workshop at Knowledge Web General Assembly 2007, 2007.

[19] Jie Li, Harold Boley, Virendrakumar C. Bhavsar, Jing Mei, "Expert Finding for eCollaboration Using FOAF with RuleML Rules", Proc. of the 2006 Montreal conference on eTechnologies, 2006.

[20] Mark Maybury, Ray D'Amore, and David House, "Awareness of Organizational Expertise", International Journal of Human-Computer Interaction, Volume 14, Issue 2 June 2002 , pages 199 - 217.

[21] Jing Zhang, Jie Tang, and Juanzi Li, "Expert Finding in A Social Network", DASFAA 2007: 1066-1069, 2007.

[22] Junichiro Mori and Mitsuru Ishizuka, "Expert Finding in Social Networks"

[23] Jian Jiao, Jun Yan, Haibei Zhao, Weiguo Fan, ExpertRank: An Expert User Ranking Algorithm in Online Communities, 2009 International Conference on New Trends in Information and Service Science, pp. 674 - 679.

AUTHORS

Ashok Bhansali completed his graduation from NIT-Jamshedpur and then pursued his M.Tech. in Computer Technology from NIT-Raipur. He is a SUN certified programmer and has more than 15 years of industrial & academic experience. He has served in various reputed organizations like Nelco, TechMahindra, SSCET etc. He is a member of ISTE, IEI, CSI and has published many research papers in international journals and conferences. He has been the chairman of various bodies and organized many conferences and STTP. Presently he is serving the OP Jindal Institute of Technology as Associate Professor in the Department of Computer Science and Engineering.

Dr. H.R. Sharma did his M.Tech Computer from Delhi University and completed his Ph.D. from IIT Delhi in Computational Mathematics. He is having more then 38 years of academic experience. He has been the chairman and member of many government and autonomous bodies. He has guided many research scholars and his research area includes Computational Mathematics, Analysis of Algorithms, AI & ES, and NLP. Presently he is working as Director CSITS Durg (C.G.). He is a member of ISTE, CSI, and IEEE.