Information Retrieval System and challenges with Dataspace

Niranjan Lal
Suresh Gyan Vihar University,
Jaipur, Rajasthan, India

Samimul Qamar
Computer Engineering
King Khalid University,
Kingdom of Saudi Arabia

Savita Shiwani
Suresh Gyan Vihar University,
Jaipur, Rajasthan, India

ABSTRACT

The advance of technology has seen increase in applications that integrate new kinds of information, such as multimedia and scientific data, unstructured, semi-structured, structured or heterogeneous data being created and stored is exploding is collectively called “Dataspace”. Data being generated from various heterogeneous sources like, digital images, audio, video , online transactions, online social media , data from sensor nodes , click streams for different domains including, retails, medical , healthcare , energy, and day to day life utilities. Information Retrieval from heterogeneous information systems is required but challenging at the same as data is stored and represented in different data models in different information systems. Information integrated from heterogeneous data sources into single data source are faced upon by major challenge of information transformation were in different formats and constraints in data transformation are used in data integration for the purpose of integrating information systems, at the same is not cost effective.

Information retrieval from heterogeneous data sources remains a challenging issue, as the number of data sources increases more intelligent retrieval techniques, focusing on information content and semantics, are required.

This paper describes the idea of Information retrieval system, Information integration which can be used in the Dataspace and heterogeneous data problems over the web.

Keywords

Information Retrieval; Dataspace; Information Retrieval Techniques; Heterogeneous Database; Dataspace

1. INTRODUCTION

1.1 Information Retrieval Introduction

Research data such as information about research results, projects, publications, organizations, researchers published on the web play more and more pervasive role in modern research. High dependence of modern research on already achieved research results produce requirements for research to have ability to retrieve research information in efficient way. Information overloading, exponential rise of amount of information makes it difficult for researcher to find relevant information. To solve these problems a number of Current Research Information Systems (CRIS) is being developed.

But in most cases such system do not solve task of providing to researcher complete and actual information with minimum information noise. Researchers are not prone to publish results about their research in information systems, publishing usually limited to researcher’s or project’s homepages. To provide actual and complete information for interested persons, information from research pages also should be included into information retrieval operations. Usually researchers ‘or policy-makers ‘demands for research information is not limited to only information stored in any one the systems. Research information in any science or technology area is scattered among a number of heterogeneous information system.

There is a strong need to gather information according request when it possible or to point researcher to systems where information can be found. It is very important to know if the gathered research information is actual and complete. So, it a necessity to find a solution for a problem data integration, which will be 1) easy to implement for any participator • flexible enough to embrace diversity and data meaning and structure in different organizations, sectors of science and states 2) powerful to go provide sophisticated information retrieval services for users One of the challenges of CRIS development is to provide access to research data which are scattered on the web pages, stored in different research information systems. The informational needs of a researcher or policymaker are very seldom limited to information from one CRIS, which usually represents research from a particular region or sector of science. There is a strong need to integrate information from different sources and to provide access to all information to users, enabling them to utilize a wide range of sources. One of the problems of such system development is heterogeneity of research information systems.

One of the main problems to deal with information managing is the weak interoperability between various databases and information systems. Especially this problem is serious when we want organize collaboration between the information systems of various departments within the organization. Data retrieval from different autonomous sources has become a hot topic during the last years. For instance, there are such data sources as employee data source, student data source, library data source etc within the same enterprise (talking of academic institution). When someone wants piece of information we need to execute n queries and possibly provide user with n such results, retrieved from n data sources. Heterogeneous data sources are searched based on user criteria and result of n sources is integrated into single source, this data source is created every time heterogeneous information systems are to be searched & structure of this single data source is dynamic and not static as such structure of this source is variable and is defined a fresh every time.

1.2 Dataspace Introduction

Designing Dataspace support platforms builds on the traditional strengths our field and will involve significant extensions of data management techniques, but it will be crucial to leverage techniques from several other fields. We mention a few here. Recent developments in the field of knowledge representation (and the Semantic Web) offer two main benefits as we try to make sense of heterogeneous collections of data in a Dataspace: simple but useful formalisms for representing ontology’s, and the concept of

URI (uniform resource identifiers) as a mechanism for referring to global constants on which there exists some agreement among multiple data providers. Similarly, as discussed earlier, several operations on Dataspace inherently involve some degree of uncertainty about the data, its lineage, correctness and completeness.

The Uncertainty in AI Community had developed several formalisms for modeling uncertainty, but these tend to be very expressive. The challenge is to find models that are useful yet simple, understandable, and scalable. Naturally, much of the data in a Dataspace will be unstructured text. Hence, incorporating techniques from Information Retrieval will play a crucial role in building DSSP. Importantly, in a complex Dataspace, users do not know exactly what they are looking for or how to interpret the results. Hence, it is important that they be able to effectively visualize results of searches and queries to better guide their exploration. Recent techniques from Information Visualization will be valuable here. Providing an effective mechanism for personal information retrieval is important for many applications, and requires different techniques than have been developed for general web search. This paper focuses on developing retrieval models and representations for personal search, and on designing evaluation frameworks that can be used to demonstrate retrieval effectiveness in a personal environment. From the retrieval model perspective, personal information can be viewed as a collection of multiple document types each of which has unique metadata. Based on this perspective, we propose a retrieval model that exploits document metadata and multi-type structure. Proposed retrieval models were found to be effective in other structured document collections, such as movies and job descriptions.

The paper is organized as follows. Section II Represents the basic of Information Retrieval System requirement with IR techniques, Section III Explains the Dataspace, Section IV Shows Heterogeneity of Data Sources and data integration and describes results already achieved in this area. Section IV describes which information model can be used in Dataspace with their challenges, and Final Section VI Concludes and future of the paper.

2. INFORMATION RETRIEVAL SYSTEM

Information retrieval is the art of demonstration, storage, organization of and access to information items. The representation and organization of information should be in such a way that the end user can access information to meet his / her information needed. In Dataspace, information retrieval finds the structured; semi-structured or unstructured data that satisfies information need from within in the Dataspace, and it inform the end user on the existence and whereabouts of data relating to his or her query.

2.1 Components of IR System

In this section we describe the components of a basic web information retrieval system shown in Fig.1, A general information retrieval functions in the following steps. 1). Crawling: The system browses the document collection and fetches documents. 2). Indexing: The system builds an index of the documents, 3). User gives the query, 4). Ranking: The system retrieves documents that are relevant to the query from the index and displays that to the user, 5). Relevance Feedback: User may give relevance feedback to the search engine. Information Retrieval Component for Dataspace responsible for Querying and Searching the system should support:1) Transition between keyword querying, browsing and structured querying , 2) Meta-data queries , 3) Queries about the source of the data, 4) Filtering and aggregation, also called monitoring of the Dataspace.

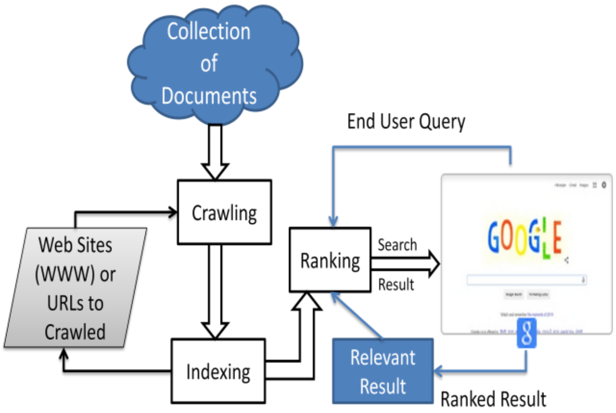

Fig.1 Components of Information Retrieval System

2.2 Why Information Retrieval

Overwhelmed by Flood of Information over the planet is shown in the Fig. 2, from the figure we can see that due to heterogeneity of data, we need different types of mechanisms to retrieve the information over the web. According to www.worldwidewebsize.com (Thursday, 08 January, 2015).there are more than, (1) 30 billion pages on the Web, (2) The Indexed Web contains 4.38 billion pages, (3) The Dutch Indexed Web contains at least 235.5 million pages. According to the www.factshunt.com (2013) there are,(1) 759 Million - Total number of websites on the Web ,(2) 510 Million - Total number of Live websites (active), (3) 103 Million - Websites added during the year i.e 2013, (4) 14.3 Trillion - WebPages, live on the Internet, (5) 48 Billion - WebPages indexed by Google.Inc,(6) 14 Billion - WebPages indexed by Microsoft's Bing, (7) CUIL.com indexed more than 120 Billion web pages,(8) Major search engines indexed at least tens of billions of web pages.

Fig. 2: Burden of Flooded Information

2.3 Information Retrieval Techniques

The goal of any information retrieval system is to satisfy user's information need. Unfortunately, characterization of user information need is not simple. User's often do not know clearly about the information need. Query is only a vague and incomplete description of the information need. Query operations like query expansion, stop word removal etc. are usually done on the query. In information retrieval, the criterion typically is the relevance of the retrieved information item with respect to the information need at hand. To achieve this goal, Information Retrieval Systems usually implement following processes:

1) In indexing process the documents are represented in summarized content form. 2) In filtering process all the stop words and common words are remove. 3) Searching is the core process of IRS. There are various techniques for retrieving documents that match with users need.

We have defined a retrieval technique [5] as a technique for comparing the query with the document representations. we can further classify retrieval techniques in terms of the characteristics of the retrieved set of documents and the representations that are used. Some techniques do not fall naturally into only a single category in this classification and other are hybrids of techniques from different categories but the scheme is useful for discussing the broad distinctions among retrieval techniques Fig. 3 gives a diagrammatic view of the classification.

The first distinction that we make among retrieval techniques is whether the set of retrieved documents contains only documents whose representation are an exact match with the query or a partial match with the query. For a partial match, the set of retrieved documents will include also those that are an exact match with the query.

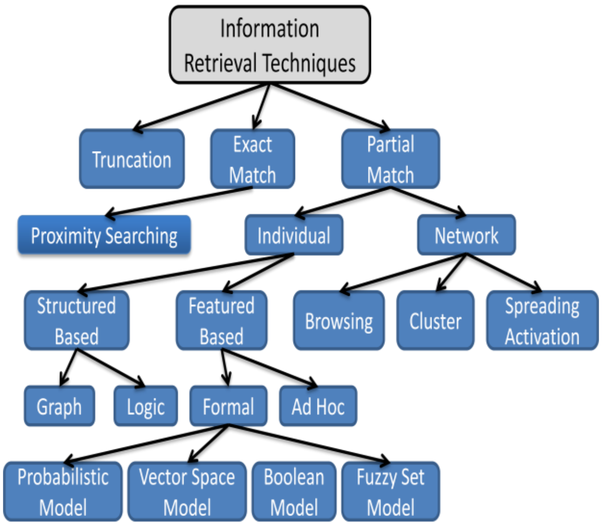

Fig.3: Information Retrieval Techniques

The next level of the classification distinguish between techniques that compare the query with individual document representative and techniques that use a representation of documents that emphasizes connections to other documents in a network. In this category, individual document are retrieved, but the retrieval is based connections to other documents and not solely on the contents of an individual documents. In the network category, we identify the subcategories of cluster based search, searches based on browsing a network of documents, and spreading activation searches.

The individual category breaks down into retrieval techniques that use a feature-based representation of query and documents and technique that use a structured based representation. In a featured–based representation, queries and documents are represented as set of features, such as index items. Features can be weighted and can represent more complex entities in the text than single words. The structured based category is divided into representation based on logic, that is those in which the meaning of queries and documents represented using some formal logic , and on representation that are similar to graph, in which documents and queries are represented by graph –like structures composed of nodes and edges connecting these nodes . Such graphs can be produced by natural language processing (e.g., semantics nets and frames) or statistical techniques.

The feature-based category includes techniques based on formal model (including the vector space model, probabilistic ranking model, fuzzy set model, and Boolean model) and the techniques based on ad-hoc similarity measures.

3. DATASPACE

As the volumes of data storage increases within and across enterprises, there is a growing need to develop efficient and effective techniques of data management. With the increase in the amount of structured, semi-structured and unstructured data available on the web as well as local data stores, the impact has been that new opportunities for using data integration technologies have been created, a general view of Dataspace is shown in the Fig 4. However, in spite of the long standing research work in data integration, this technology seems to have had a limited impact in practice.

To a large extent, data integration mechanisms are manually coded and tightly bound to specific applications. The limited adoption of data integration technology is partly due to its cost-ineffectiveness [1]. The problem of data integration has been investigated for a relatively long period of time spanning about two decades with the aim of providing end users with a transparent access to data sets that reside in multiple data sources and are stored using heterogeneous representations. Data integration has numerous potential applications, e.g., it can be used for providing cross-querying of data stored in databases that belong to multiple departments or organizations, or to enhance collaboration in large scientific projects by providing investigators with a means for querying and combining results produced by multiple research labs [1].

More precisely, the specification of schema mappings (in such a way that, data structured under the source schemas is transformed into a form that is compatible with the integration schema against which user queries are issued) has been found to be both time and resource consuming, and has also been determined as a critical bottleneck to the large scale deployment of data integration systems [2]. Dataspace is a current technique of managing data. Since it's envisioned in 2005, Dataspace has growingly been fronted as the new technique of data integration [3,11]. Data integration is an important research topic since it aims at providing transparent access to data that is stored in various data repositories that are often using different underlying data models.

3.1 Why Dataspace?

Dataspace are clearly distinguished from traditional data integration approaches due to the fact that they provide for integration on a pay-as-you-go fashion. This way it is cheaper. More importantly, Dataspace do not require upfront effort for semantic integration; they focus on data co-existence instead. Additionally, Dataspace provide higher degrees of scalability due to the perceived nature of the entity relationships [1], [2], [3], [4].

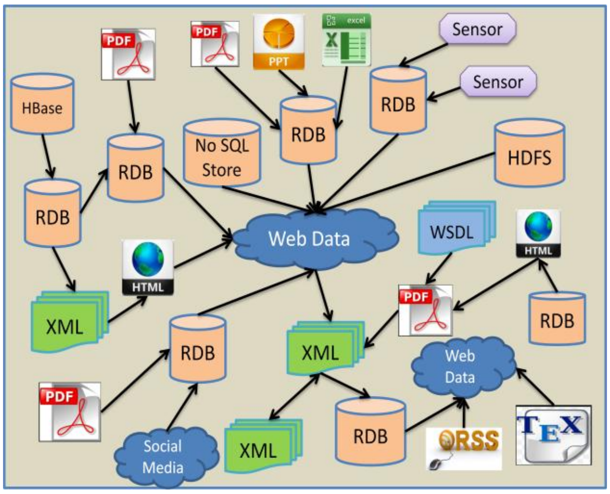

Fig. 4 A View of Dataspace

Examples

One example of a Dataspace described as a personal Dataspace is described by [2] as a personal Dataspace in which users access data stored in a set of personal data repositories.

That set of repositories may include private file systems that currently exist on a user's desktop as well as private emails of a particular user. In the case of personal Dataspace just like the case of other Dataspace, users often experience difficulties understanding which items spread across their repositories/sources are related to each other in the same context. Although users are likely to search their data sources with search engines, the results obtained by these systems are not enriched with contextual information. Users may typically desire to access the various versions of a certain file that exist in their Dataspace, view files as well as emails worked on approximately the same time, or extract emails in the same project of a certain document.

A second example of a Dataspace which is available as a public Dataspace is called Google Base. It is further described as a very large, self-describing, semi-structured, heterogeneous database. Google Base certainly consists of a set of tuples with attribute values. Each tuple entry Te is regarded to consist of a number of attributes with matching values. An illustration of a Dataspace tuple is shown in Fig. 5.

T_e{(name:GalaxyPrime),(Color:black),(mane: Samsung), (tel: 8756),(addrs:Plotno ,8, Alwar,Raj.),(website :Samsung.com)}

Fig. 5 An Illustration of Dataspace Tuple

In this case, the tuple in the Dataspace set can be considered as a tuple in an existing Dataspace. In this second example, the data set is enormously sparse due to the heterogeneity of data, which are continuously contributed by users around the world.

4. HETEROGENEITY OF DATA SOURCES

In addition to the decentralization, the effectiveness of information retrieval is further worsened by the variety of heterogeneity present in the data sources. In each of these sources, data is organized using a different system (operating systems, SQL Vendor Implementations etc), based on different conceptual models, and on different formats "system level heterogeneity" is considered to be nowadays much easier than before (e.g. via ODBC/JDBC connections on relational databases), much interest is laid on the so called "semantic heterogeneity", which appears every time there is a more than one way to structure a body of data. Semantic heterogeneity seems to be an unavoidable burden in data sharing and manipulation, since people tend to model their data according to their own understanding of the reality. This of course is fundamentally different for each individual. In that sense, heterogeneity is to be found in data models, conceptual schemas, and of course the mind of the users. Page Layout [9].

4.1 Solution

The basic principle of data integration is to combine (integrate) selected information sources from specific domain, in a way that a whole new data Source is generated. The end-user, when querying for data, has the illusion of interacting with one single system, which presents him a unified logical view of the data available. The first attempts to address information integration issues in enterprises where based primarily on data warehousing techniques. Traditional solutions prescribe creation of new data source on Information integration from heterogeneous data sources, which is not cost effective.

5. INFORMATION RETRIEVAL IN DATASPACE

5.1 Why Information Retrieval in Dataspace

Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure. In IR systems, the information is not structured, it is the collection of heterogeneous data i.e. Data from any number of sources, largely unknown, unlimited in many varying formats or unknown contents and partially unstructured i.e. no pre-defined data model, usually text (e.g. text documents, Word documents, Email messages, RSS feeds, Audio files, Video files), semi-structured (e.g. XML, LATEX, web data, Electronic Data Interchange(EDI), scientific data), or structured i.e. Pre-defined and machine readable, a locatable, sometimes relational 'data model' usually of real-world objects (e.g. Relational database (RDB), Customer Relationship Management(CRM), Enterprise Resource Planning(ERP) data, collectively called Dataspace.

5.2 Information Retrieval models for Dataspace

The basic question in Dataspace information retrieval is that How to find data of interest in within a collection of heterogeneous data?

The fundamental IR models can be classified into Boolean, vector, and probabilistic model [6, 7,10]. These IR model are well suited for the Dataspace information retrieval.

The most used data model in information retrieval are : the Vector Space Model(VSM) representing resources and queries as vectors in Dataspace ; the Probabilistic Ranking Model(PRP) using the notion resources pertinence with respect to a specific query; and the Inference Network defined as an hybrid of VSM and PRP models.

1) Boolean model: The Boolean model allows for the use of operators of Boolean algebra, AND, OR and NOT, for query formulation, but has one major disadvantage: a Boolean system is not able to rank the returned list of documents. In the Boolean model, a document is associated with a set of keywords. Queries are also expressions of keywords separated by AND, OR, or NOT/BUT. The retrieval function in this model treats a document as either relevant or irrelevant [7].

2) Vector Space Model: The vector space model can best be characterized by its attempt to rank documents by the similarity between the query and each document .In the Vector Space Model(VSM), documents and query are represent as a Vector and the angle between the two vectors are computed using the similarity cosine function. Similarity Cosine function can be defined as:

Where,

$$Sim(\dot{d}_J, \dot{q}) = \frac{\dot{d}_J.\dot{q}}{\|d_j\|\|q\|} = \frac{\sum_{i=1}^{N} w\dot{_{i,J}} w\dot{_{i,q}}}{\sqrt{\sum_{i=1}^{N} \dot{w}_{i,J}^2}\sqrt{\sum_{i=1}^{N} \dot{w}_{i,q}^2}}$$

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

$$\dot{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

Vector Space Model have been introduce term weight scheme known as if-idf weighting. These weights have a term frequency (tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (idf) factor measuring the inverse of the number of documents that contain a query or document term [8].

3) Probabilistic model: The most important characteristic of the probabilistic model is its attempt to rank documents by their probability of relevance given a query. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query, or not. Instead of probabilities, the probabilistic model uses odds O(R), where O(R) = P(R)/1 − P(R), R means ‟document is relevant‟ and ¯ R means ‟document is not relevant‟ [8].

5.3 Information Retrieval challenges for Dataspace

There exists several information retrieval techniques, discussed above that can be applied within our context in order to process the documents of the Dataspace and determine the appropriate data contained in the data sources.

Challenges faced by search engines, in Dataspace to retrieve the relevant data or information.

Challenge 1: One particular challenge is the large scale, given by the huge number of WebPages available on the Internet (for example, about 30 billion WebPages were indexed by Google in 2005).

Challenge 2: Another challenge is inherent to any information retrieval system that deals with text: the ambiguity of the natural language (English or other languages) that makes it difficult to have perfect matches between documents and user queries.

Challenge 3: Designing DSSPs will involve significant extensions of data management techniques, but it will be crucial to leverage techniques from several other fields.

Challenge 4: Several operations on Dataspace inherently involve some degree of uncertainty about the data, its lineage, correctness and completeness

Challenge 5: Incorporating techniques from Information Retrieval will play a crucial role in building DSSP.

Challenge 6: Recent techniques from Information Visualization will be valuable.

6. CONCLUSION AND FUTURE

Due to that inference abilities and schema exploration can make development of Research Information System more easy then conventional technologies like Relational Database management systems because exploration of domain knowledge is very crucial for CRIS systems. Data Integration was considered to be “an area of intellectual curiosity “at its early years, the advent of information sharing nowadays is calling for effective integration approaches realized in practice. Users are not compromising with low standards of information accuracy and are willing to find the right information at the right time. The research community, thus far, has shown excellent progress in dealing with the most crucial problems presented on the way of integrating data, however, further challenges arise constantly:

The expansion of semi & unstructured data (XML) for example implies that data sources are even more complex and difficult to handle. Coping with semantic heterogeneity in such scenarios seems almost impossible. However, research is getting even more intense and promising ideas are expected to develop. We can choose one of the models for Information Retrieval models for Dataspace as discussed above.

There are several directions for future work in this area .It would be interesting to develop new appropriate information retrieval model under the framework which can be used for in Dataspace to produced better query results, and new direction to do research in Information Retrieval and Dataspace challenges.

7. ACKNOWLEDGMENT

We would like to acknowledge the support and input from our family, Dean Research, Faculty members of Suresh Gyan Vihar University Jaipur. Dean and colleagues from Mody University of Science & Technology Laxmangarh, Sikar (Raj.) India.

8. REFERENCES

[1] K. Belhajjame,N. Paton,S. Embury, Feedback-Based Annotation, Selection and Refinement of Schema Mappings for Dataspace, ACM EDBT 2010, March 22–26, 2010, Lausanne, Switzerland, 2010.

[2] M. Franklin, A. Halevy, and D. Maier, “Principles of Dataspace systems”, Proc. of Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2006), ACM Press, p. 1-9., ISBN 1-59593-318-2.

[3] M. Franklin, A. Halevy, D. Maier, From databases to Dataspace: A new abstraction for information management, ACM SIGMOD Record 34 (4) , 2005, 27-33.

[4] P. Ziegler, K. Dittrich, Data Integration — Problems, Approaches, and Perspectives, Springer, Berlin Heidelberg, 2007 Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] J. Belkin, Nicholas J., and W. Bruce Croft " Retrieval Techniques 'Volume 22,1987.

[6] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", ACM Press, ISBN: 0-201-39829-X.

[7] Anwar A. Alhenshiri, Web Information Retrieval and Search Engines Techniques, Al- Satil journal,PP: 55-92.

[8] G. Salton and M.J. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill 1983,

[9] Er. Majid Zaman,, et al " Information Integration for Heterogeneous Data Sources" IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 640-643.

[10] Shibwabo Bernard Kasamani ,, Ateya Ismail Lukandu " A Survey of the State of Dataspace " International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 02– Issue 06, November 2013.

[11] Niranjan Lal, Samimul Qamar, "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application (ICACEA),pp 565-572, March 2015.

9. AUTHOR PROFILE

Mr. Niranjan Lal received B.E. Degree in Information Technology from Rajasthan University, India, in 2005, and M.Tech Degree in Information Technology from Guru Gobind Singh Indraprastha University Delhi, India in 2007. He a Research Scholar at Suresh Gyan Vihar University Jaipur, Rajasthan Indian, and He is currently Assistant Professor in Dept. of Computer Science & Engineering at Mody University of Science & Technology Lakshmangarh, Sikar , Rajasthan, INDIA . His research areas are Computer Networks, Network Security, and Wireless Sensor networks, Cloud Computing, Dataspace, Mobile Computing, and Android Application Development.

Dr..Shamimul Qamar , He is B.Sc, in 1992 from, Meerut University, Meerut, India. B.Tech (ECE) in 1996 from MMMEC, Gorakhpur University, India. M.Tech (Information System) in 2000 from AMU, Aligarh, India, and Ph.D (Computer Science & Engineering) in 2006 with Grade A from Indian Institute of Technology Roorkee, India. (Ph.D Thesis Title "Capacity & Quality of Service for CDMA Communication Network With Diversity & Handoff").He currently Professor in Computer Network Engineering Department at King Khalid University, Abha, Kingdom of Saudi Arabia. His research areas are Networking, Cloud Computing and Database.