Fuzzification of Web Objects: A Semantic Web Mining Approach

Tasawar Hussain[1], Muhammad Abdul Qadir[2] and Sohail Asghar[3]

**[1] Department of Computer Science, Mohammad Ali Jinnah University
Islamabad, 44000, Pakistan**

**[2] Department of Computer Science, Mohammad Ali Jinnah University
Islamabad, 44000, Pakistan**

**[3] University Institute Information Technology, University of Arid Agricultural
Rawalpindi, 44000, Pakistan**

Abstract

Web Mining is becoming essential to support the web administrators and web users in multi-ways such as *information retrieval; website performance management; web personalization; web marketing* and *website designing*. Due to uncontrolled exponential growth in web data, knowledge base retrieval has become a very challenging task. The one viable solution to the problem is the merging of conventional web mining with semantic web technologies. This merging process will be more beneficial to web users by reducing the search space and by providing information that is more relevant. Key web objects play significant role in this process. The extraction of key web objects from a website is a challenging task. In this paper, we have proposed a framework, which extracts the key web objects from web log file and apply a semantic web to mine actionable intelligence. This proposed framework can be applied to non-semantic web for the extraction of key web objects. We also have defined an objective function to calculate key web object from user's perspective. We named this function as *key web object* function. KWO function helps to fuzzify the extracted key web objects into three categories as *Most Interested, Interested,* and *Least Interested*. Fuzzification of web objects helps us to accommodate the uncertainty among the web objects of being user attractive. We also have validated the proposed scheme with the help of a case study.

Keywords: *Semantic Web Mining, Web Mining, Key web Objects, Website Ontology, Web Log File, Object Objective Function, Fuzzification.*

1. Introduction

Today's world is behaving like a global village due to massive growth of World Wide Web (WWW). The internet has revolutionized the modern world and connected the people under one umbrella (WWW) (Antoniou and Harmelen, 2003). According to Allemang and Hendler (2008) web is an open source for everyone under the slogan of AAA (Anybody can say Anything about Any topic). This openness is main driving force for the explosive growth of Internet. At the same time, web is facing the major drawback of unstructured and unformatted web contents. Due to this drawback, relevant information retrieval from web is becoming a challenging job of today's informative world.

Each website is composed of web pages and each web page consists of a number of objects (Stumme et al., 2006) and these objects could have different formats such as text, audio, video, animation and images (Vela´squez et al., 2011a). In a semantically enriched website, these objects represent the defined structure of web contents of the webpage with properly defined metadata. Moreover, in case of traditional web, web contents are without proper structure and format. Web objects contain the web contents of a website, for which users are looking for. Whereas the key web objects are the web contents of the given webpage that have sufficient user support. Vela´squez et al. (2011a) defined the key web objects as group of objects that attract the users. Consequently, key web objects are not only captivating its users but also represent the structured web contents.

How a web object is interested from user point of view is difficult to judge. There is no proper mechanism on websites to get the user feedback about the contents of website. Only web log files are the important source to study the user behavior and interests on website. There are three main sources of web log files such as web server log file; proxy server log file; and client web log file (Hussain et al., 2010c). All the three sources have their own pros and cons but server web log file is considered more authentic and is widely used in to digging out the user behavior and interests. Primarily, web log files are used to monitor the performance of server but not to get the user feedback etc. To study the users' behavior from web log

file requires the techniques, which may be applied on web log file known as web usage mining (WUM). The major and serious drawback of web log file is to record the WebPages traversed by users in a given session. Log file has no information regarding the web objects. Owing to this drawback, the extraction of key web object is a challenging job.

With Unstructured data and without common vocabulary the web is an uncontrolled horse. Machines are third important pillar in the web as illustrated in Fig. 1 but in conventional web, machines are inactive and only play the role of dormant partner. As a result, these are some of the common features of the traditional web, which are core obstacles to retrieve the information. Consequently, the keyword based search engines provide most of the irrelevant stuff. For that reason, to cart, to the point information from the web, requires the expertise of users and a lot of time.

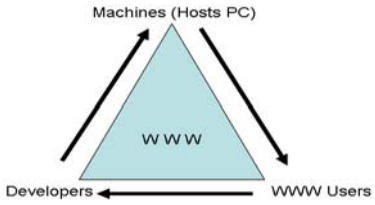

Fig. 1 Pillars of www

Consequently, such kinds of major drawbacks of the traditional web have created the gaps and opportunities for innovation. Tim Burner Lee, the founding father of the Web, also coined the idea of semantic web as a solution to these problems. Tim Burner Lee not only provided the layered framework of the semantic web but also pointed some of the main features of the semantic web(Stumme et al., 2006).

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

The eventual solution to the traditional web is semantic web. We have to equip the web contents with organized structure and in machine processable form to breach the gap between machines and its users to accomplish the tasks by applying techniques that are more intelligent.

As the core objective is to provide knowledge and information to human beings in a swift way by adding intelligence to machines. Provision of a structure to the web data is one aspect of the solution, another equally important aspect is the development of techniques to find hidden intelligence in the data. Consequently, we are combining the semantic web techniques and data mining techniques to equip the human being with proper knowledge extraction (Stumme et al., 2006) (Fig. 2).

Fig. 2 Semantic web mining

The rest of paper is organized as follows: Section II gives details on existing literature on key web objects extraction. Section III is about the proposed semantic web mining technique to extract the key web objects. In section IV, we evaluated the proposed scheme with the help of a case study. Section V concludes the paper and presents the future research direction.

2. Literature Review

The extraction of key web objects is an offshoot of semantic web and web usage mining. Whereas, Stumme et al. (2006b) amalgamated the both semantic web and web usage mining techniques in the form semantic web mining. For the extraction of key web objects, web mining approaches such as content mining; structure mining; and usage mining (Vela´squez et al., 2011b) are in common practice by researchers. A handful literature is available on semantic web mining approaches. We have reviewed the literature in two directions semantic web approaches and web mining approaches.

2.1 Extraction Web Objects based on Semantic Web

Semantic web has provided the structured to the web resources into machine understandable form and the extraction of web objects by applying semantic web mining techniques is an observable fact. Vela´squez et al. (2011a) applied semantic web technique and web usage mining techniques to extract key web objects. In their proposed research, they developed the ontology of the given website to enrich the web contents semantically. Most of the tasks were performed manually. For the extraction of key web objects from log file, they trained the users and give prior knowledge about the web objects

and their importance to website users. Trained users do not give a picture of the true usage of website.

Nie Z., et al. (2006) defined the web object as the basic data unit about which we gather web information, indexed and ranked them. Web objects behave like concepts such as conferences, papers, authors etc. These objects provide the web information to the web users; Attributes are defined as properties to define the web objects. The attributes are divided into three categories such as Key Attributes; Important Attributes; and Other Attributes. Authors used object level information extraction approach to extract web objects (Fig. 3). Object blocks are the containers of web object on a given web page of a website and grouped.

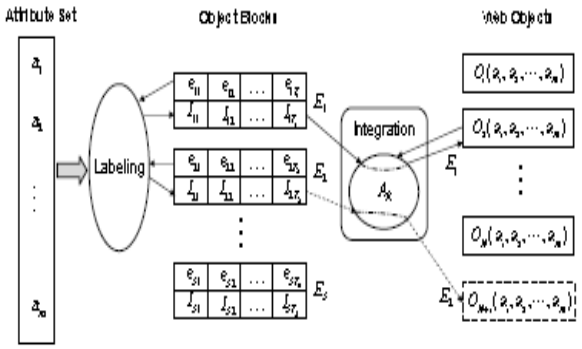

Fig. 3 Web object extraction (Nie et al., 2006)

The object extraction process is not semantically enriched. Website ontology was not built while attributes and concepts were defined. The website designer categorized web objects while the web objects are not given weightage from users' point of view.

Miao et al. (2009) extracted the Data Records (Data Objects) from the website by applying the tag path clustering. Data Records were extracted in three steps such as *Detecting Visual Information; Data Record detection;* and *Semantic Level Nesting Detection.* Fig. 4 gives relationship between ancestor node and descendant node within a cluster.

The triplet < p_i, S_i, O_i > represents the visual signal and p_i represents the tag path; S_i visual signal vector; and O_i occurrence. The detection of repeating visual signal vector is termed as clustering problem and for clustering the similarity measure is key concept for clustering technique. To find the occurrence O_i of visual signal S_i in a cluster C_i the ancestor /descendant relation between visual signal vectors is found. In Semantic –Level Nesting Detection step, extracted data records are organized in semantic categories.

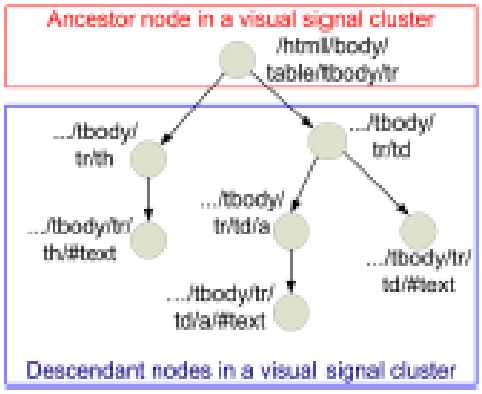

Fig. 4 Maximal ancestor containing one data record (Miao et al., 2009)

2.2 Extraction of Web Objects via Web Mining

Primarily web mining is composed of three main categories such as Web Structure Mining; Web Content Mining; and Web Usage Mining (Hussain et al., 2010b, Pokorny and Smizansky, 2005). Web structure mining is used to discover the association between web pages within website or from one website to another. The structure mining helps to a search engine to pull data relating to a search query directly to the linking web page from the website. Web structure mining (WSM) focuses two main challenges of the web such as irrelevant search and indexing of web pages. Web Content Mining (WCM) is the mining of text, images, videos, audios and other web data available on web. With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query. Web Usage Mining (WUM) exploits the web user behavior. Web log file records all the users interactions with website and web server is responsible to maintain the web log file. Log file can also be maintained at proxy server and at client site as well. However, in literature, it has been observed that web server log file is more consistent to extract user behavior.

According to Liu (2005), large amount of web information is available in structured data objects and to identify these data objects two new approaches were introduced. These approaches are Wrapper Induction and Automatic Extraction. In Wrapper Induction, pages of website are taken and then labeled manually. After labeling any suitable data mining technique is applied to extract rules or patterns. In second approach, single page is taken to

extract the patterns and each single page is consists of multiple data objects.

Pokorny and Smizansky (2005) proposed a page relevance ranking based on the page content exploration. The importance of web page is marked based on contents available in the form of web and importance of a term is specified with respect to a given user query *q* and it is based on its statistical and linguistic features. To calculate the page content relevance an aggregation function (Eq 1) is used.

$$Sec_{moment}(s) = \sum_{i=1}^{m} \frac{w_i^n}{n} \quad (1)$$

On the basis of this aggregation function, important terms on a given web page are marked.

As we have reviewed different techniques to identify the web objects. Major drawback in literature review is lack of proper framework to manage the traditional and semantic web at a same time. Similarly pruning of web objects is also common practice. Time is an important factor to mark the key web objects which has been totally ignored in literature available. Our proposed research not only jacketing the both web technologies but also introduces fuzzifcation of the web objects rather than pruning.

3. Proposed Methodology

For the fuzzification of web objects, we have to combine the two emerging technologies; *Semantic Web* and *Web Usage Mining (WUM)*. Semantic approach enriches the website with proper and structured format of web objects. WUM helps to find out the page frequency and average time of each visited page from web log sessionization. WUM gives the judgment of website users to locate the key web objects. The Key web object (KWO) Function is applied on each web object. KWO calculates the values of each web object based on information provided by the web log usage. Subsequently web objects are fuzzified by applying the Mamdani Fuzzy Model. And we are able to fuzzify the web objects as *Least Interesting*; *Interesting*; and *Most Interesting*. In Fig. 5, a proposed framework has been given to fuzzify the web objects. In following sections, we will discuss the phases of proposed framework.

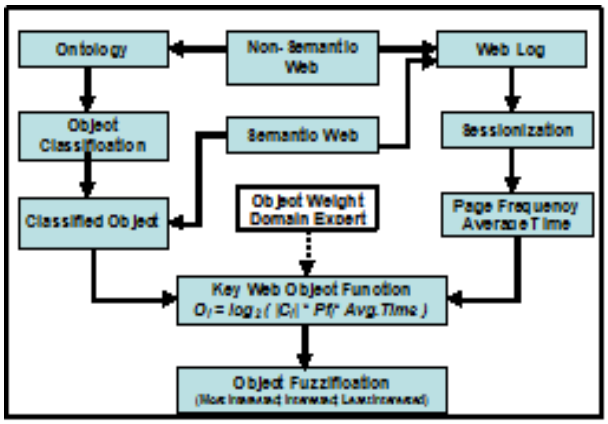

Fig. 5 Proposed framework

3.1 Classification of Web Objects

Most of the websites are in non semantic and the main objective of applying semantic web approaches is to provide the structure to the non semantic world. For a non semantic web, ontology (Fig. 6 & Fig. 7) of a given website is created by applying Website Parse Template (WPT) (Manukyan et al., 2009). Each web object is defined by one or more concepts. If an object is defined by more than one concept, then such web objects are more likely to be key web objects. For calculating the KWO function, numbers of concepts defining an object play pivotal role. In object classification step, we find out the number of concepts per web objects. For semantic web, we have build-in ontology of given website, and we required to classify the web concepts per web object.

3.2 Web Log Sessionization

In second phase, we take the web log file of the given website. After preprocessing the web log, we find out the different user sessions. For sessionization, we adapted the methodology given by (Hussain et al., 2010b). After sessionization, we find out the frequency of each visited page by different users and average time consumed by users on a page of website. Page frequency gives knowledge about the importance of page contents. Larger the count of page frequency indicates the worth of page contents. On the other hand, time spent by each user on a single page of also indicates the importance of page contents from user point of view. Different user may spent different time on the same page due to their own interests on different web contents of the page. For assigning equal weight to each user, we take the average time spent by each user.

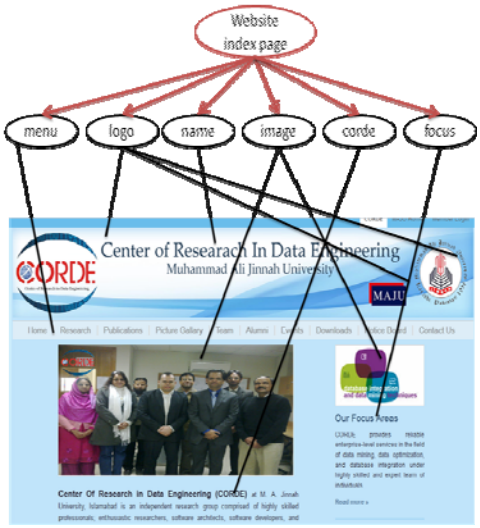

Fig. 6 Website ontology

```
<ontology name="Given Website">
    <concept name="menu">
        <has object="home"></has>
        <has object="Research"></has>
        <has object="Publication"></has>
        <has object="Pictures"></has>
        <has object="Team"></has>
        <has object="Alumni"></has>
        <has object="Events"></has>
        <has object="Download"></has>
        <has object="ContactUs"></has>
    </concept>
    <concept name="Logo">
        <has object="logo1"></has>
        <has object="logo2"></has>
        <has object="logo3"></has>
    </concept>
    <concept name="name"></concept>
    <concept name="image">
        <has object="image1"></has>
        <has object="image2"></has>
    </concept>
    <concept name="corde"></concept>
    <concept name="focus"></concept>
</ontology>
```

Fig. 7 XML code of website ontology

3.3 Key web objectFunction (KWO)

In this step, we combine the both the semantic web and web usage mining to calculate the key web object function of each web object. From semantic web, we have web concepts count for each web object and at the same time, web usage mining provides us the page frequency and average time consumed by different users. The following Eq. 2 calculates objective function for each web object.

$$Q_i = \log_2(|C_i| * |P_{ij}| * Avg_{ij}) \quad (2)$$

Where C_i be the number of concepts of *ith* web object, P_{ij} be the frequency of *ith* web page, which contains ith web object, and Avg_{ij} be the average time of *ith* page consumed by users.

3.4 Fuzzification of Web Objects

The objective function values of each web object are taken in fuzzy format and Mamdani Fuzzy Model is applied to obtain the fuzzy model of web objects. We divided the fuzzy region into three fuzzy member functions such as; *Least Interesting; Interesting; and Most Interesting*. If an object has higher value of KWO, it is the most interesting web object from user viewpoint and from website concept hierarchy. In least interesting membership function, we put the objective functions values from 0 – 4. For interesting and most interesting membership functions, we assign the objective function values from 2 – 8 and 6 – onward.

4. Case Study

We take the example in which we have ten objects of a given website. Each web object is defined by one or more web concepts. Table 1. gives details about the web objects and number of concepts from which a particular web object is inherited or defined.

Table 1. Concepts Containment of Objects

Objects	No of Concepts
O_1	1
O_2	2
O_3	5
O_4	3
O_5	5
O_6	2
O_7	4
O_8	3
O_9	4
O_{10}	1

In next step we mark the web objects to the relevant web page by the help of domain expert. Table 2. gives details about the web pages and their relevant objects.

Table 2. Objects on Web Pages

Page	Objects
P_1	O_1, O_2, O_3
P_2	O_4, O_5, O_6

P_3	O_7, O_8
P_4	O_9, O_{10}

In next step, we take the web log file of the website, after preprocessing the web log; we count the frequency of each page traversed by users in different users' sessions. Similarly, we calculated the average time per page as time consumed by different users while navigating website. Table 3 & 4 give details about the page frequency and average time of per page.

Table 3. Web Log Page Traversal Frequency

Page	Frequency
P_1	5
P_2	3
P_3	4
P_4	2

Table 4. Average Time Per Page

Page	Avg. Time
P_1	3
P_2	2
P_3	3
P_4	4

After calculating page frequency and average time of each page, we calculated the key web object function of each object. Table 5 gives the key web object values.

Table 5. Key web object Function

Objects	Values
O_1	3.91
O_2	4.91
O_3	6.23
O_4	4.17
O_5	4.91
O_6	3.58
O_7	5.58
O_8	5.17
O_9	5.00
O_{10}	3.00

The fuzzification of web objects is given in following Fig. 8a & 8b. For web object fuzzification, we applied the matlab fuzzy box. For 0-4 key web object function values, we mark the least interesting objects, for the values 2- 8, we have interesting web objects and for 6 to onward key web object values, we have most interesting web objects.

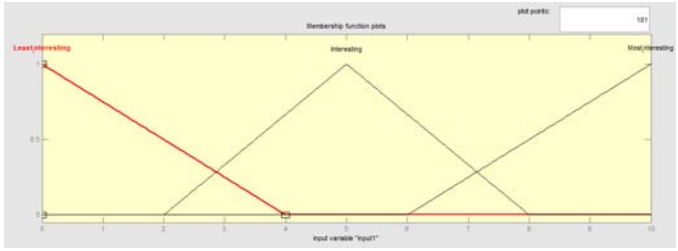

Fig. 8a Fuzzification of web objects

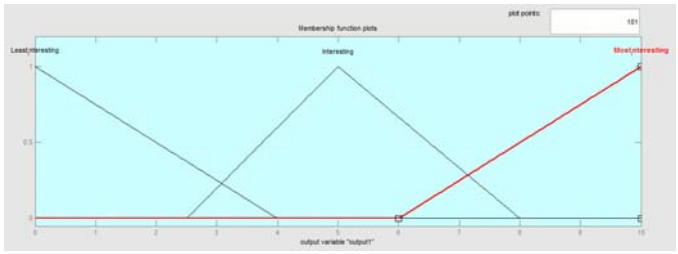

Fig. 8b Fuzzification of web objects

5. Conclusion

In this research paper, we have discussed the various techniques to extract the key web objects by applying semantic web mining approaches. We also proposed a framework to fuzzify the web objects. The major drawback is to pick the true behavior of users about the web objects from log file. Log file gives no knowledge about the web objects. It assigns equal weight to each web object present on a web page. In future, we required a technique, which can portray the true object level user behavior from users' aspects. Domain experts' weight age scheme can also be utilized for the better identification of web objects from users' point of view.

Our propose framework is novel approach to fuzzify the web objects which can produce better results different web applications such as Information Retrieval; e Advertising; and Recommended Systems. We applied the case study to the proposed framework. Our approach is a milestone in considering the all web objects of a given website. We are not pruning the key web objects based on threshold.

References

[1] ALLEMANG, D. & HENDLER, J. 2008. *Semantic Web for the Working Ontologist*, Morgan Kaufmann Publishers is an imprint of

Elsevier. 225 Wyman Street, Waltham, MA 02451, USA.
[2] ANTONIOU, G. & HARMELEN, F. V. 2003. *A Semantic Web Primer*, The MIT Press Cambridge, Massachusetts London, England.
[3] CHAMBERS, N., ALLEN, J., GALESCU, L., JUNG, H. & TAYSOM, W. Year. Using Semantics to IdentifyWeb Objects. *In:* Proceedings of the National Conference on Artificial Intelligence, 2006 USA. 6.
[4] HUSSAIN, T., ASGHAR, D. S. & FONG, S. 2010a. A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining. *6th International Conference on Advanced Information Management and Service (IMS)*. Seoul, Korea.
[5] HUSSAIN, T., ASGHAR, D. S. & MASOOD, D. N. 2010b. Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence. *2010 6th International Conference on Emerging Technologies (ICET)*. Islamabad.
[6] HUSSAIN, T., ASGHAR, D. S. & MASOOD, D. N. 2010c. Web Usage Mining: A Survey on Preprocessing of Web Log File. *Information and Emerging Technologies (ICIET), 2010* Karachi, Pakistan: IEEE.
[7] LIU, B. Year. Web Content Mining. *In:* The 14th International World Wide Web Conference (WWW-2005), May 10-14, 2005 2005 Chiba, Japan.
[8] MAEDCHE, A. & STAAB, S. 2001. Ontology Learning for the Semantic Web. *IEEE Inteligent system*, 16**,** 72-79.
[9] MANUKYAN, A., MANILYAN, A. & SAYADYAN, A. 2009. *Website Parse Template* [Online]. Available: www.w3c.org [Accessed].
[10] MIAO, G., TATEMURA, J., HSIUNG, W.-P., SAWIRES, A. & MOSER, L. E. Year. Extracting Data Records from the Web Using Tag Path Clustering. *In:* International World Wide Web Conference Committee (IW3C2), 2009. ACM 978-1-60558-487-4/09/04, 981-990.
[11] NIE, Z., WU, F., WEN, J.-R. & MA, W.-Y. 2006. Extracting Objects from the Web. *In:* REPORT, T. (ed.) *ICDE, MSR-TR-2004.* Microsoft Research.
[12] POKORNY, J. & SMIZANSKY, J. 2005. Page Content Rank: An Approach to the Web Content Mining. In proceedings of IADIS International Conference Applied Computing, Algarve, Portugal, 2005
[13] STUMME, G., HOTHO, A. & BERENDT, B. Year. Usage Mining for and on the Semantic Web. *In:* Data Mining: Next Generation Challenges and Future Directions. AAAI/MIT Press, Menlo Park,CA, 2004 CA. 467–486.
[14] STUMME, G., HOTHO, A. & BERENDT, B. 2006. Semantic Web Mining State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4**,** 124–143.
[15] STUMME, G. & MAEDCHE, A. Year. FCA-MERGE: Bottom-Up Merging of Ontologies. *In:* 7th Intl. Conf.on Artificial Intelligence (IJCAI '01). , 2001 Seattle, WA. 225–230.
[16] VELA´SQUEZ, J. D., DUJOVNE, L. E. & L'HUILLIER, G. 2011. Extracting significant Website Key Objects: A Semantic Web mining approach. *Engineering Applications of Artificial Intelligence***,** 10.

Mr. Tasawar Hussain received his MS(CS) degree from the Department of Computer Sciences, Muhammad Ali Jinnah University, Islamabad, Pakistan in 2010. Previously, he obtained the M.Sc degree in Mathematics from Peshawar University, Peshawar, Pakistan in 1995. Currently, Mr. Hussain is a PhD (CS) scholar at Muhammad Ali Jinnah University, Islamabad, Pakistan. His research activities are in the areas of data mining, web mining, semantic web mining and constraint base mining.

Dr Muhammad Abdul Qadir is a professor in Faculty of Engineering & Sciences at Mohammad Ali Jinnah University, Islamabad, Pakistan. He is the head of the CDSC (Center for Distributed & Semantic Computing) research group http://www.cdsc.jinnah.edu.pk). Dr. M.A. Qadir received the MSc degree in electronics from Quaid-i-Azam University Islamabad, Pakistan, and the PhD degree from University of Surrey, UK in parallel processing/distributed computing. His research activities are in the areas of grid computing, parallel processing, distributed computing and context aware computing. Dr. M.A. Qadir is a member of IEEE.

Dr. Sohail. Asghar is a Director at University Institute Information Technology, University of Arid Agriculture, Rawalpindi, Pakistan. Previously he was Associate Professor in Department of Computer Sciences at the Mohammad Ali Jinnah University, Islamabad, Pakistan and was Assistant Professor of Information Technology and Head of R&D, Faculty of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan. Formerly he was Research Associate and Assistant Lecturer in Clayton School of Information Technology, Faculty of Information Technology at Monash University, Melbourne, Australia. In 1994, he graduated with honors in Computer Science from the University of Wales, United Kingdom. From 1994 to 2002, he worked as a Senior Software Engineer in a software company in Islamabad. He then obtained his PhD in Information Technology at Monash University, Melbourne Australia in 2006. http://sohailasghar.wordpress.com