I.J. Information Technology and Computer Science, **2013, 03, 49-55**
Published Online February 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2013.03.06

A Novel and Efficient Method for Protecting Internet Usage from Unauthorized Access Using Map Reduce

P. Srinivasa Rao
MVGRCE, India
psr.sri@gmail.com

Dr. **K. Thammi Reddy**
GITAM, India
thammireddy@gitam.edu

Dr. **MHM. Krishna Prasad**
JNTUK, India

***Abstract*—** The massive increases in data have paved a path for distributed computing, which in turn can reduce the data processing time. Though there are various approaches in distributed computing, Hadoop is one of the most efficient among the existing ones. Hadoop consists of different elements out of which Map Reduce is a scalable tool that enables to process a huge data in parallel. We proposed a Novel and Efficient User Profile Characterization under distributed environment. In this frame work the network anomalies are detected by using Hadoop Map Reduce technique. The experimental results clearly show that the proposed technique shows better performance.

***Index Terms*—** Mapreduce; Hadoop; Distributed Computing

I. Introduction

Today, we're drowning in data i.e. People upload videos, take pictures on their cell phones, text friends, update their Face book status, leave comments around the web, click on ads, and so forth. The exponential growth of data gave us an opportunity to know more about the hidden secrets by which one can make golden nuggets. Existing techniques were becoming inadequate to process such large datasets. Google was the first to publicize Map Reduce –a system they had used to scale data processing needs[9]. This system aroused a lot of interest because many other businesses were facing similar scaling challenges, and it wasn't feasible for everyone to reinvent their proprietary software. Doug cutting [7] saw an opportunity and led the change to develop an open source version of the Map Reduce system called Hadoop. Soon after Yahoo and others rallied around to support this effort Today, Hadoop is a core part of the computing infrastructure for many web companies such as, Yahoo, Facebook, LinkedIn and Twitter. Many more traditional businesses, such as media and telecom, are beginning to adopt this system tool.

Hadoop is a versatile framework, which allows new users to access the power of distributed computing. By using distributed storage and transferring code instead of data. Hadoop avoids the costly transmission step when working with large datasets. Moreover, the redundancy of data allows Hadoop to recover from a single node fail. The most prominent and well-supported ones have officially become subprojects under the umbrella of Apache Hadoop project. These subprojects include

- Pig – A high-level data flow language
- Hive – A SQL-like data warehouse infrastructure
- HBase – A distributed, column-oriented database modeled after Google's Bigtable
- ZooKeeper – A reliable coordination system for managing shared state between distributed applications
- Chukwa – A data collection system for managing large distributed systems

Among all Hadoop HDFS is more powerful because of its powerful feature such as open source and have a fault tolerance because of Secondary Name Node which will replicate the data that will be processed by Hadoop.

This paper presents a realization of extraction of user profile method, which adopts Hadoop framework, and

MapReduce distributed programming technology, to mine the user's interests from implicitly mining web log content. We used the network flow data from the web server of an Educational Institution, after partitioning the data into equal sizes then we given these portioned files as input to Hadoop to make clusters based on their IP addresses to identify different anomalies in the data set.

The rest of this paper is organized as follows. In section 2 the related work is discussed and in section 3 the proposed system Model and architecture was presented. In section 4 the approach and in section 5 analyses is presented. Finally, conclusions are made in section 6.

II. Related Work

Hadoop is a software platform which is easy for development and processing mass data [7]. It is written in Java. Hadoop is scalable, economical, efficient and reliable. It can be deployed to a big cluster composed of hundreds of low-cost machines. The user's programs written with the Hadoop API can be executed in parallel to improve efficiency. The programmers can only care the main operation regard less of all relevant details of distribution. Map Reduce [8] is an important technology of Google. It is a kind of programming models, which comes from the traditional functional programming ideas to deal with mass data [9]. HDFS divides a large data file into several blocks stored in different nodes. However the Map Reduce, which is constructed above HDFS, copies the user's tasks to the several nodes in cluster for parallel executing. All nodes in the Hadoop are distributed to different machines in cluster, and communicate with each other to run the entire framework [10]. kyong-Ha Lee [1] provided a survey of Parallel Data Processing with Map Reduce. Kumar [4] presented how to identify Network Anomalies Using clustering Techniques in Weblog Data which provided a way to analyze log file of proxy server. Ken Mann [13] studied and provided Distributed computing with Linuxand Hadoop. Arun C Murhty[14] given an overview on Next Generation of Apache Hadoop Map Reduce-The Scheduler where he explained background knowledge of resource utilization by Hadoop. Kashyap Santoki [15] explained his view on indexing and searching on a Hadoop Distributed File System from which knowledge of Hadoop HDFS can be undersood. Spiros Papadimitriou [11] written a paper on DisCo: Distributed Co-clustering with Map-Reduce A Case Study Towards Petabyte-Scale End-to-End Mining. Rajiv Gupta [16] gave his explanation on Efficiently Querying Archived Data using Hadoop. Michael Cardosa[17] explained how to Explore MapReduce Efficiency with Highly-Distributed Data.

III. Proposed System Model

Today, the massive data that is geographically distributed has derived a need for distributed computing to reduce the time consumed to process the data. Also, when the size of the datasets extend beyond the capacity of a single storage, if is extended to distribute them in multiple independent computers [1]. Hadoop is treated as most powerful frame work which uses resources that are geographically distributed. The key feature of Hadoop is its faulttolarence feature. The key element of Hadoop is HDFS (Hadoop Distributed File System)

Hadoop [3] is a java-based software frame work which enables the data intensive application in a distributed environment [2]. The key feature of the Hadoop is that as there is a chance of the computing elements and the storage to fail it maintains multiple copies of working data to ensure that the processing can be redistributed to the failed nodes. Hadoop is made up of different elements and at its bottom, the Hadoop Distributed File System (HDFS) which stores the files across the storage nodes in the Hadoop cluster and above the HDFS is the Map Reduce engine.

Map Reduce is a scalable and fault-tolerant data processing tool which enables to process a massive volume of data in parallel with many low-end computing nodes and is popularized by Google. Map Reduce is the parallel programming model that consists of two phases viz., Map phase and Reduce phase. The Map Reduce is simple and efficient for computing aggregate. Thus, it is often compared with "filtering then group-by aggregation" query processing in a DBMS.

Now-a-days, speed, complexity and size of the network is growing rapidly, and the networks became open to public access.For knowing access patterns users like students, teaching , nonteaching and others will access internet.

To verify the works of Firewalls the users will use Proxyserver and Virutal tunnels.

Users also access the net for by passing the load.Hence there is a tremendous increase in number and type of intrusions so that making it impossible for human analysis. To make analysis simple we can use different techniques for network analysis[4].

We are in this experiment using Hadoop Map Reduce technique to partition the weblog data to detect the network anomalies. In applying visualization to internet security, researchers exploit the human ability to process visual information quickly enables the complex task of network security monitoring and intrusion detection to be performed accurately and efficiently as discussed in [5][6].

3.1 Algorithm

Step 1: Read input log file;

Step 2: if log file too large split into parts;

Step 3: Process with Map Reduce;

Step 4: retrieve output of each terminal in Hadoop;

Step 5: analyze log file;

3.2 Architecture

HDFS divides a large data file into several blocks stored in different nodes. However the Map Reduce which is constructed above HDFS, copies the user's tasks to the several nodes in a cluster for parallel executing as shown in Figure 1.

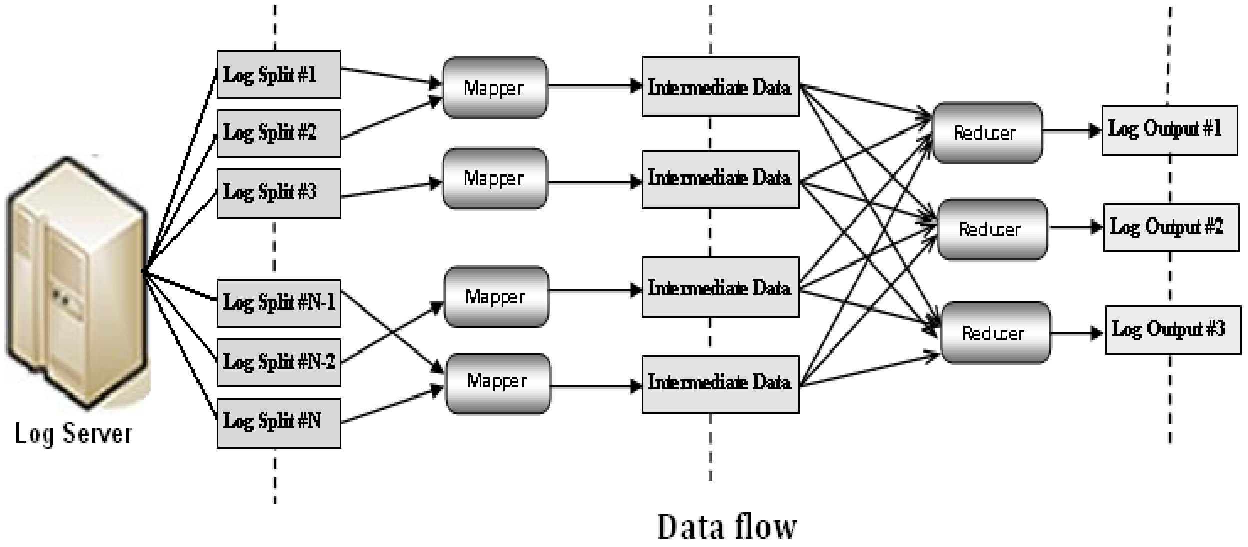

Fig. 1: Map Reduce Architecture

As shown above in our proposed system model , all users will login at remote terminal with their IP address and do their transactions. All these transactions will be recorded in a log file of Proxy server that will be used as input to our developing system. The input file will be partitioned as it is very large to process and is given to Hadoop Map Reduce for further processing.

Figure 2 shows how log file will be created at server and how the log file is processed using hadoop is also shown.

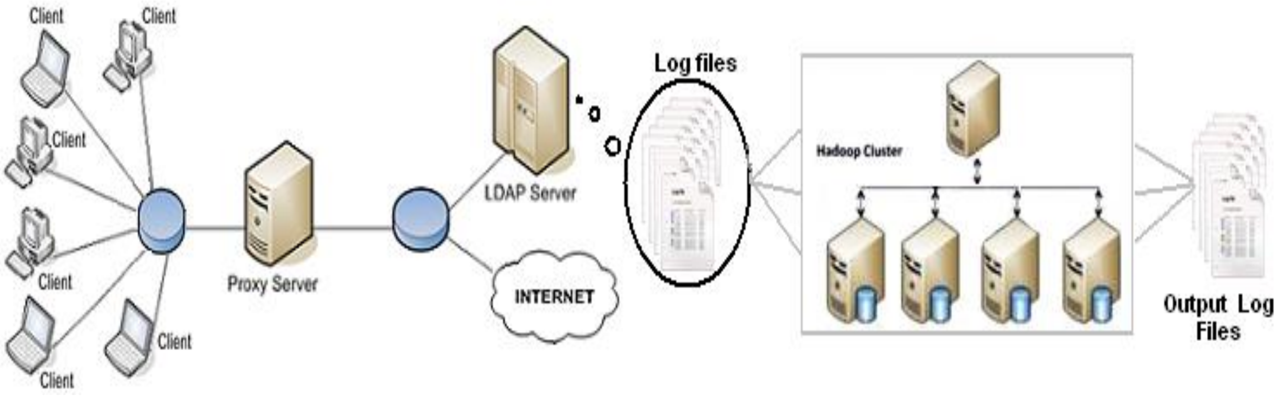

Fig. 2: Creation of log file at Proxy Server

All nodes in the Hadoop are distributed to different machines in cluster and communicate with each other to run the entire frame work as shown in Figure 3.

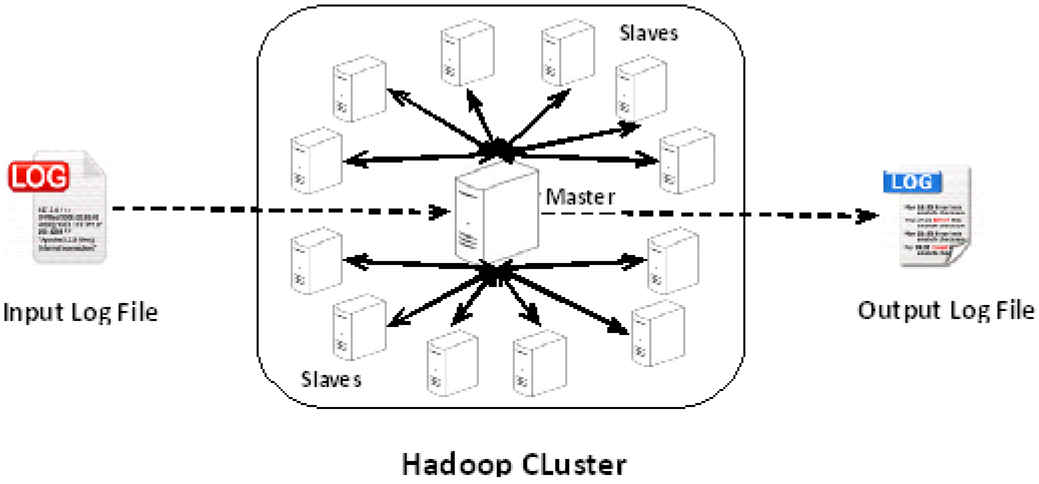

Fig. 3: Hadoop Cluster

IV. Experimental Methodology

In this paper we demonstrate the process of extracting the user behavior by considering the IP address and URL pattern of log file from the server with a capacity of 4GB and processing the log file to filter the data. The approach used for filtering the data is Map Reduce. Map Reduce is an essential distributed grep-sort-aggregate (a distributed executed engine) for select project through sequential scan and then partitioned using hashing and sort-merge group by [11].

The Map Reduce is capable of processing many terabytes of data on huge number of machines by splitting the input data sets into independent blocks which are dependent on number of nodes and the size of data. In practice Map Reduce is a 2 phase process: the Map phase and the Reduce phase.

The Map phase maps one set of data to another set of data by a one-to-one rule. The Reduce phase reduces the set of data which are found to be redundant by many-to-one rule. It is model which is encapsulated inside the Hadoop framework[10].

As the web content is increasing dramatically, it is highly difficult to aim at the user's profile. In such cases, personalized services[12] such as personalized information retrieval, personalized website and personalized recommendation systems play a vital role. These services extract the user's profile. There are 3 methods for extracting the user's profile: server mining, customer offered initiatively and systematic learning independently. Among these existing methods, we in this paper have used server mining technique. Because every server maintains the user's access log record, it is easy to assess the user's character by analyzing this log file. To achieve this process, the log file in the server which is about 4GB size is utilized for experimentation. It is highly difficult to read that huge file without use of DFS hence we have partitioned further it into various blocks of files have been fed as input to the Map Reduce function.

In our experimentation, we have configured 4 node with core2 duel processor, 2GB RAM, 320 GB HDDs and the software used are Hadoop 0.20.0 version with ubunto 10.4.0.

The information about the user's behavior is analyzed by extracting the user's details such as IP address and URL. The data format is as follows.

Table 1: Sample Log File Format

```
902351618.864  440 120.65.1.1  TCP_MISS/304
110 GET
http://www.webtrends.com:8005/Images/search.gif
- DIRECT/www.webtrends.com  -902351639.632

120.65.1.1  TCP_REFRESH_MISS/200  2014  GET
http://www.webtrends.com/news/news.txt -
DIRECT/newsroom.compuserve.com
text/plain902351659.571  274

207.65.46.33  TCP_CLIENT_REFRESH/200  1467
GET
http://www.webtrends.com/news/headlines.txt - -
DIRECT/www.webtrends.com
text/plain902351691.541

120.65.1.1  TCP_REFRESH_MISS/200  2014  GET
http://www.webtrends.com/support/support.txt -
DIRECT/www.webtrends.com
text/plain902351708.872  286

207.65.46.33  TCP_MISS/200  1384  GET
http://www.webtrends.com/news/headlines.txt - -
DIRECT/www.webtrends.com  text/plain
```

After processing a weblog data file such as shown above the following is sampe output format that we can obtain after it is processed by hadoop MapReduce.

Table 3: Sample output Format

IP Address	URL's
192.168.5.0	http://www.w3schools.com
192.168.5.1	www.google.com,http://apache.techartifact.com
192.168.5.2	http://englishinlearning.blogspot.in/
192.168.5.3	http://www.youtube.com
192.168.7.5	http://www.facebook.com,www.yahoo.com,www.gmail.com

Due to the bulk data, we have used distributed programming model which is based on Hadoop and run on Hadoop clusters deployed by 4 machines. To compare the results, we have run it on even single machine also.

V. Results analysis & Performance Evaluation

To compare the experimental results, initially the program is run on single machine with Map Reduce and cluster deployed by four machines with Map Reduce node.

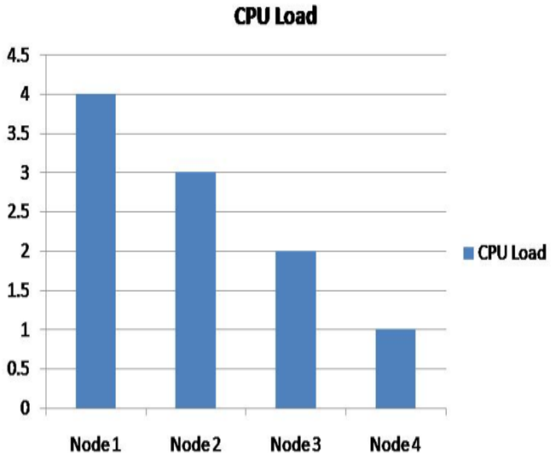

Fig. 4: CPU Load

As shown in Figure 4 when number of nodes are increasing the load on CPU is gradually decreasing.

Table 2: CPU Load

No. of Nodes	Time in Sec.
1	416.879
2	386.324
3	257.476
4	126.458

Table 2 shows how time will be reduced as number of nodes are increasing.

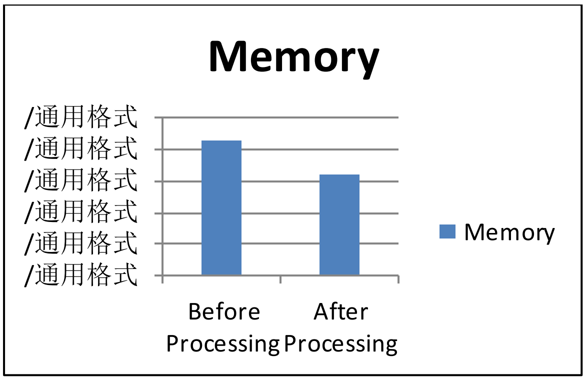

Fig. 5: Memory utilization

As shown in Figure 5 the memory required for storing log file before processing is more when compared to the memory required after processing. Because of the removing of redundant IP addresses (Millions of lines) in the log file the memory utilized will be drastically reduced by this experiment.

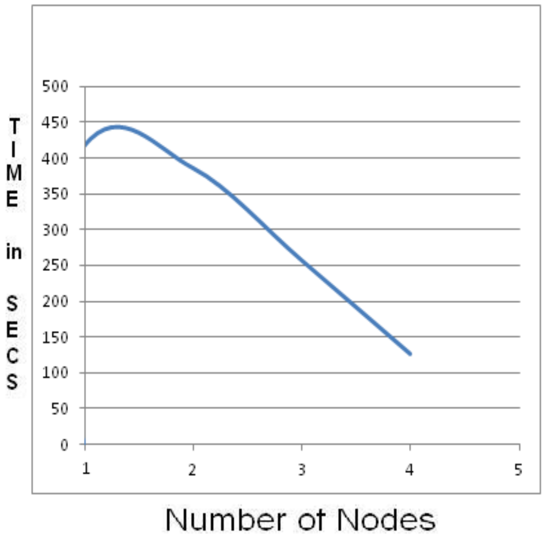

Fig. 6: Time versus No.of Nodes

As shown in Figure 6. If we increase number of number of node (clients), time required to process the data is decreasing.

VI. Conclusion

With the increase in large amounts of internet data every day it is becoming highly complex to process the bulk data in less time. To encounter this difficulty, the Google along with Apache, has introduced a parallel programming model Map Reduce on Hadoop which enables the parallel batch processing.

In this paper, we have shown the single Cluster Map Reduce architecture suitable for situations when data and computer resources are widely distributed. We experimented on the data of 4GB size and compared the results with single node, 2 nodes and 4 nodes. It is observed that the performance increases as the number of nodes increases. Hence the Map Reduce will reduce the time taken to process a huge file, eliminate redundancies in the log files for analysis, reduced memory utilization and due to the parallel programming model, the CPU load is also distributed. As a result of the lessons learned from our experimental evaluations, we have provided recommendations for when to apply the various Map Reduce architectures, as a function of several key parameters: workloads, network topology and data transfer costs and data partitioning. Therefore our experiments yield better results when compared with others.

References

[1] Kyong-Ha Lee, Hyunsik Choi, Bongki Moon "Parallel data processing with MapReduce: A Survey" December 2011.

[2] Ping ZHOU, Jingsheng LEI, Wenjun YE"Large-Scale Data Sets Clustering Based on Map Reduce and Hadoop", Binary information Press, 1553-9105, December 2011.

[3] "Hadoop", http://hadoop.apache.org

[4] B. Kiran Kumar, A. Bhaskar "Identifying Network Anomalies Using Clustering Technique in Weblog Data" ,ijca,Volume 2 No.3, June 2012.

[5] Kai Shuang1, Yin Y ang "X-RIME: HADOOP-BASED LARGE-SCALE SOCIAL NETWORK ANALYSIS" Proceedings of IC-BNMT20 10

[6] Kulsoom Abdullah, Chris Lee, Gregory Conti, John A.Copeland. "Visualizing Network Data for Intrusion Detection", Proceedings of the IEEE 2002.

[7] DougCutting,"HadoopOverview", http://research.yahoo.com/node/2116

[8] Jeffy Dean,Sanjay Ghemawat. MapReduce, "Simplified Data Processing on Large Clusters", OSDI04: Sixth Symposium on Operating System Design and Implemention, Ssn Francisco,CA,December, 2004.

[9] Jeff Dean. "Handling Large Dataset at Google: Current System and Future Direction", http://labs.google.com/people/jeff

[10] HUANG Lan*, WANG Xiao-wei, ZHAI Yan-dong, YANG Bin "Extraction of User Profile Based on the Hadoop Framework" IEEE 2009.

[11] Spiros Papadimitriou Jimeng Sun "DisCo: Distributed Co-clustering with Map-Reduce A Case Study Towards Petabyte-Scale End-to-End Mining" ICDM.2008.142.IEEE 2008.

[12] LIU Ni-na, "Research on the Web Mining and Personalized Search Engine". Master degree theses of Zhejiang University,2005.

[13] Ken Mann " and provided Distributed computing with Linux and Hadoop. "Freelance March 2011.

[14] Arun C Murhtygiven " Next Generation of Apache Hadoop Map Reduce-The Scheduler where he explained background knowledge of resource utilization by Hadoop." Freelance March 2011.

[15] Kashyap Santoki "indexing and searching on a Hadoop Distributed File System from which knowledge of Hadoop HDFS can be undersood." July 2010.

[16] Rajiv Gupta "Efficiently Querying Archived Data using Hadoop",CIKM'10,ACM978-5/10.

[17] Michael Cardosa "Exploring MapReduce Efficiency with Highly-Distributed Data".MapReduce'11,ACM, USA June 2011.

[18] Jiong Xie "improving MapReduce Performance through Data placement in Hetrogeneous Hadoop Clusters".IEEE 2010.

Authors' Profile

P.Srinivasa Rao, currently working as Sr.Asst.Professor in CSE of MVGR College of Engineering. He is having Over 08 years of teaching experience. His research includes Data warehousing and Mining, Distributed Computing, Image Processing etc.

Dr. K. Thammi Reddy, currently working as the Director of Internal Quality Assurance Cell (IQAC) and Professor of CSE. at Gandhi Institute of Technology (GITAM) University, Visakhapatnam. He is having vast experience In teaching, Research,Curriculam Design and

consultancy. His research areas include Data warehousing and Mining, Distributed computing, etc.

Dr. MHM.Krishna Prasad, currently working as Associate Professor in IT Dept. of JNTUCE,Vizianagaram. He is having vast experience In teaching, Research and Curriculam Design. His research areas include Data warehousing and Mining, Computer Networks,Embedded Systems, etc.

How to cite this paper: P. Srinivasa Rao, K. Thammi Reddy, MHM. Krishna Prasad,"A Novel and Efficient Method for Protecting Internet Usage from Unauthorized Access Using Map Reduce", International Journal of Information Technology and Computer Science(IJITCS), vol.5, no.3, pp.49-55, 2013.DOI: 10.5815/ijitcs.2013.03.06