ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

MapReduce based Method for User Friendly XML Keyword Search

Samaha Sadique K, Swaraj K P

M. Tech Student, Dept. of Computer Science and Engineering, Govt. Engineering College, Thrissur, Kerala, India

Assistant Professor, Dept. of Computer Science and Engineering, Govt. Engineering College, Thrissur, Kerala, India

ABSTRACT: XML data is platform independent. So it is used for information exchange in many applications. Various systems were proposed to find XML data based on a given keyword. Some systems focused on query based search and some focused on searching XML without the use of any query. Kapase and Shinde proposed a user friendly technique that uses a combination of LCA and SLCA method for XML keyword search. But this technique is not fast enough in the case of large amount of data. To address this limitation, MapReduce technique is incorporated with LCA and ELCA method for fast and user friendly keyword search of XML data. Parallel algorithms are used to process plenty of XML documents in Hadoop environment. A distributed LCA and ELCA computation method is designed, where each net node computes LCA and ELCA independently and only a little information needs to be transmitted.

KEYWORDS: Keyword search, XML, Distributed Computation

I. INTRODUCTION

XML data is independent of platform. So it is widely used.
There is a need to search XML data based on keyword for many applications.Researchers developed many techniques for XML searching. Some used query languages like XPath and XQuery to find XML data and others proposed XML keyword based search methods without the use of any query. In the former approach, a user creates a XML query, submits it to the system for processing, and retrieves relevant answers from XML data. This method requires knowledge about XML data structure and data content. If the user has only a little knowledge about XML query languages then he/she will face the difficulty of adopting this approach. Sometimes user will require trying few possible keywords; this could be a tedious and time consuming process.The latter approach is better because of its user friendly nature. Users don't need a good knowledge on XML query languages. User just needs to type the required keyword and the result is retrieved in the form of XML data based file.

In this paper, a combination of two keyword based search techniques called LCA(Lowest Common Ancestor) and ELCA(Exclusive Lowest Common Ancestor) methods are proposed for user friendly keyword search. However, the fast growth of data day by day brings big challenge of processing data.To overcome this problem we collaborate above keyword search method with MapReduce technique for fast and efficient user friendly keyword based XML search. In this paper, we study the problems of various keyword based search methods .The main challenge in keyword search is the presence of user query with multiple keywords. This type of multiple keyword query must be answered correctly. To achieve our goal, we study top-kalgorithm to answer keyword queries in XML data.

We study the effective ranking function and earlytermination techniques. In this paper we proposed techniqueto improve the response time or search speed of keyword search on XML data by using MapReduce along with an efficient search technique like LCA and ELCA.The proposed system therefore makes XML search fast as well as user friendly.This proposed prototype will be effective andfeasible to design and develop real world based XML datasearch application.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

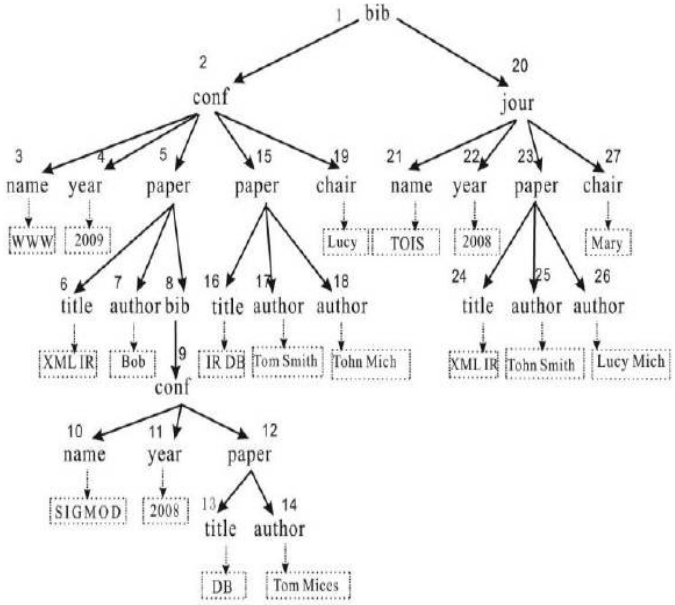

Fig 1:Parent-Child based XML Tree Structure

Figure 1 shows Parent-Child XML tree structure which contains content nodes. Content nodes are the nodes which contain the keyword to be searched. For example, content nodes of keyword "DB" are 13and 16.This structure is considered for keyword based search.

Our contribution can be summarized as follows:

1. We applied LCA and ELCA based search methods for user friendly keyword search on XML data.
2. We used a novel ranking scheme for XML keyword search.
3. We collaborated MapReduce technique with the above mentioned search methods for fast and efficient search in the case of large XML data.

The rest of this paper is organized as follows: Section 2 presents the related work; Section 3 presents an overview of proposed system; Section 4 presents the details of the System model and we conclude this paper in Section 5.

II. RELATED WORK

Bast and Weber [4] proposed a complete search intextual document based on "type less find more" technique which allows query keyword appear at any place in the document to find relevant answers. However, this techniquedoes not allow minor error between query keyword and result.Xu and Papakonstantinou proposed another technique whichuses *LCA based method* for XML keyword search. Manyalgorithms for use the notation ofLCA for XML keyword search [3].But this technique sometimes returns irrelevant and poor quality results.Also they use the "AND" semantic between keywords and ignore the answers that contain some partial part of keywords and also need to find candidates first before ranking them, and this approach alone is not efficient for computing the results.

To overcome the problems of LCA approach another technique called *Exclusive LowestCommon Ancestor* (ELCA) method is used which helps the users to find more reasonable results. J Chen and Lyad A. Kanjb showed how top-*k* algorithm works on XML database [9]. G Li, Chen Li, J Feng and L Zhou defined how to retrieve particular keyword present in XML file and how to retrieve accurately if particular keyword is not perfectly matching [8]. Kapase and

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

Shinde proposed a method for ranking which uses two ranking functions to compute rank or score between node *n* and keyword k_i [1].

MapReduce based XML search was studied based on the contribution of Zhang, Li and Liu [2].They proposed a kind of parallel XML keyword search algorithm and realized on a MapReduce programming model. MapReduce consists of two functions called Map and Reduce which do fast and distributed processing of massive files.

III. PROPOSED SYSTEM

The proposed system is an improvement totechniques introduced in [1].The main motive of proposedsystem is to perform user friendly XML keyword search in a distributed and fast way. Users can perform XML search without having any prior knowledge about the XML query languages.So a non-database expert can easily obtain the results based on input keyword. The main advantage of proposed system is its userfriendlyinterface design. This new system makes use offorward index structure approach in order to obtain top-*k*results with less processing time and populate the results.Additional advantage of this system is that it supportsapproximate keywords search even in the presence of minor errors.For example,keywords *'Hello', 'ello', 'Hllo', 'Helo'*will return the sameresult.

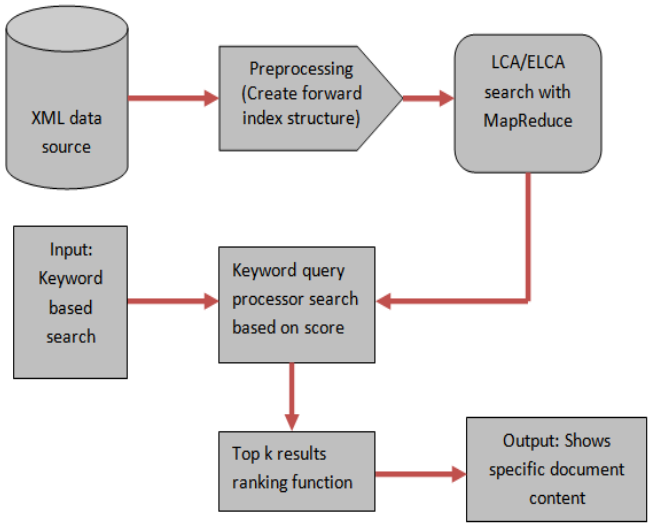

Fig 2: Proposed System Architecture

Figure 2 shows the architecture of the proposed system. The proposed system uses forward indexing concept thatcreates forward index trie structure inthe pre-processing step. Thisforward indexinghelps to improve keyword based queryprocessing. System takes input from the user interface. Input can be in theform of single or multiple keywords. Here datasource is XML data files (i.e. text files containing XML data) .Pre-processing step generates a forward index trie structure which is then searched using LCA/ELCA based method. Since this computation takes lot of time MapReduce framework is being incorporated with LCA/ELCA approach for fast and efficient searching. Results obtained after the search is ordered with the help of ranking function. Rankingfunction assigns numeric values to obtain results andgenerates top-*k* results based on ranking score using trie-based index structure. Based on ranking scoreresults are filtered and displayed. User will select the required document node and result is populated.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

IV. DETAILS OF THE SYSTEM

In this section we showed the details of the proposed methods. Major problem associated with standard XML query processing tool is with their syntax which requires XML query languages like XPath and XQuery to find XML data. These approaches need a good knowledge of XML query languages which does not go well with a non-database expert user.To address these problems a user friendly approaches are introduced here.They are LCA and ELCA based search method [7] [20]. Since data is being growing day by day the LCA/ELCA approach alone cannot withstand users' needs. So a distributed and parallel MapReduce framework is combined with this technique for fast searching. Top-k results are retrieved based on the ranking score after searching. The techniques used in this paper are described in detail in the following section.

A. LCA BASED METHOD

The *Lowest common ancestor* (LCA) is a concept in data structure graph theory. Let *T be* a rooted tree with n number of nodes. The LCA between two nodes aand*b* is defined as the lowest node in *T* that has both *a* and *b*as descendants. The LCA of *a* and *b* in *T* can also be defined as the sharedancestor between *a* and *b* that is located farthest from theroot. This is the most commonly used method to search XMLdata. The forward index trie structure contains content nodes.*Content nodes* are parent node ofthe keyword. For e.g. consider keyword *"DB"* to besearched in XML file shown in Figure 1. Contentnodes of "DB" are 13 and 16.Another data structure that is being used here is the *Inverted List*. Inverted List consists of leaf nodes which has all nodes that contain keyword.

Procedure:

1. The content nodeand its inverted list are retrieved from the XML data based on the given keyword.
2. Lowest common ancestor of content nodein the inverted list is identified.
3. The rooted sub tree at LCA is returned as the answer of the given keyword.

For example: user gave "www db" as input query then*content nodes* of *"db"* are {13, 16} and for *"www"* are {3}.The LCA identified of this *content node* are{12,15,2,1}.Here if we see that nodes {3,13,12,15} are more appropriate answers compared to nodes {2,1}.This shows that LCA does not find relevant and high quality results always. This is the main limitation of LCA based approach.

B. ELCA BASED METHOD

To overcome the demerits of LCA approach, ELCA approach is introduced which finds the exclusive LCA from the index structure.It states that an LCA method isELCA if it is still LCA after excluding its descendants LCA. For example: User typed keyword as *"db, tom"*then the *content node* for *"db"* are {13, 16} and for*"tom"* is{14, 17}. Identified LCA of the *content node* are{2, 12, 15, 1}. Here ELCA is {12, 15}. Rooted sub treeis displayed which are appropriate answers of given query.Node 2 is not an ELCA as it's not LCA after excludingnodes {12, 15}.

C. MAPREDUCE FRAMEWORK

Hadoop [10] is a flexible virtual grid operating system architecture which performs the storage and processing of big data.It has two components to perform these functions.They are Hadoop Distributed File System (HDFS) and MapReduce. HDFS is used for reliable storage and MapReduce is used for processing big data.Hadoop is capable of handling structured,semi-structured as well as unstructured data.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

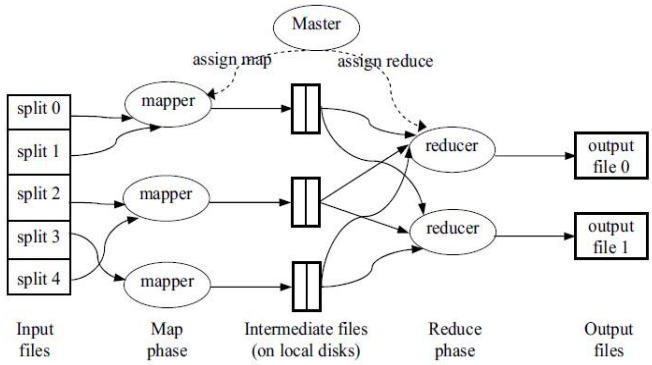

Fig 3: MapReduce Architecture

Figure 3 shows the architecture of MapReduce framework. MapReduce has a master-slave architecture where master node assigns tasks to slave nodes. It is the computation part of Hadoop.It consists of two functions called map and reduce.Both are user defined functions.Input to mapper will be in the form of key/value pairs and map function will convert this key/value pairs into intermediate key/value pairs.Then these intermediate key/value pairs are sorted and grouped based on the key and returns a list of intermediate key/value pairs with values having identical keys as a single entry.This list of intermediate key/value pair is given to reducer.The reduce function in the reducer will convert thisinput to a list of values and this is the final output.

In this paper, MapReduce is used along with ELCA search technique. To implement this technique the Parent-Child XML tree is represented in the form of Dewey tree. Each node in the tree is represented using Dewey code.Dewey Coding System is a system of classifying books and other works into ten main classes of knowledge with further subdivision in these classes by the use of numbers of a decimal system. Figure 4 shows the Dewey XML tree in which root node *bib* has a Dewey code 0.Its children will have Dewey code will be 0.0, 0.1 and so on.A MapReduce algorithm is proposed here for fastandefficient XML keyword search.

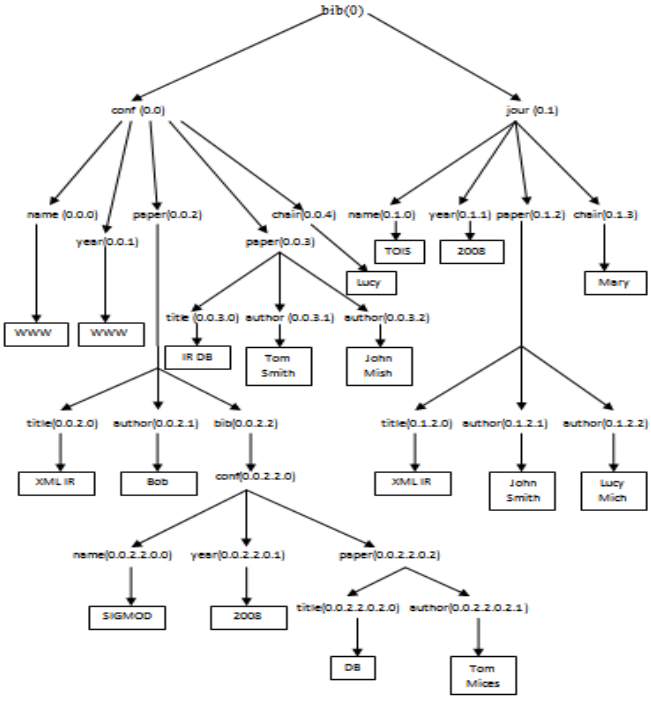

Figure 4: Dewey XML Tree

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

Three algorithms are used in this paper. Algorithm 1 first selects the keywords to be searched and its corresponding Dewey Code, and saves all Dewey code of same keyword to the same set.

Algorithm1. SelectMapReduce

```
1:SelectMapper(key=offset, value=(dewey, keyword)){
2: if keyword= w_i then
3:   output (keyword, dewey);
4: end if
5:}
6:SelectReduce(key=(keyword, dewey), values){
7: for all the same keyword do
8:   put dewey into S_i
9: end for
10: output (keyword, S_i);
11:}
```

Algorithm 2 arranges the Dewey code of each keyword in the ascending order.

Algorithm2. SortMapReduce

```
1:SortMapper(key=keyword, value=S_i){
2: length =the size of dewey in S_i;
3: output(length, value);
4:}
5:SortReducer(key=length, values){
6: according to length, sort S_i from small to large
7: output((keyword, S_i), Text());
8:}
```

Algorithm 3 realizes the parallel version of ELCA method.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

Algorithm 3: ELCA MapReduce

```
1: ELCAMapper (key=offset, value= (keyword, Si) {
2:   for i=2 to m do
3:     Put Si into Ss;
4:   end for
5:   Split S1 into several Sets;
6:   for each Set in S do
7:     result= ELCA (Set, Ss);
8:     if result is not null then
9:       output (size of result, result);
10:    end if
11:  end for}
12:  ELCAReduce (key= (size of result, result), values) {
13:    output (key, values);
14: }
```

D. *RANKING FUNCTION*

Two ranking functions are there to compute rank or score between node *n* and keyword k_i. Two cases are considered in this paper:

i. Node *n* contains keyword k_i.
ii. *Node n* doesn't contain keyword *ki* but has descendant containing *ki.*

First case:The relevance/score of node *n* and keyword *ki* is computed by the equation

$$SCORE_1(n, k_i) = \frac{ln\left(1 + tf\left(k_{i,n}\right)\right) * ln\left(idf\left(k_i\right)\right)}{(1-s) + s * ntl(n)}$$

: number of occurrences of *ki* in sub tree rooted *n.*

: ratio of number nodes *n* in XML data tonumber of nodes contain keyword *ki.*

: length of node *n* /*nmax* length.

s :Constant set to 0.2

nmax: node with maximum number of terms

Second case: Ranking function to compute the scorebetween *n* and *ki* is derived by the following equation:

$$SCORE_i(n, k_i) = \sum \alpha^{\rho(n,p)} * SCORE_i(p, k_i)$$

p:set of pivotal nodes.

α:constant set to 0.8.

$\rho(n,p)$:Distance betweennode n and pivotal node p.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

E. TOP-K RESULT ALGORITHM

The trie based index structure is used to compute the results.For computing top-*k* answers heap based method is used which uses the partial virtual inverted lists which contain the higher scorenodes.

Procedure:

1. Sort scores in the inverted lists.
2. Construct MAX heap, such that node contain <*node,score*>form.
3. The top node of MAX heap is the highest score of node and is deleted with heap adjusted.
4. Deleted node with score value <= *T* (threshold) are taken into result set and return result set if top-*k* answers are retrieved.

V. CONCLUSION

A lot of techniques were proposed by the researchers for XML keyword search. Some requires the user to be familiar with XML query languages which is a tedious task for a non-database expert user whereas in some approaches there is no need to have good knowledge in query languages. One of the user friendly methods for XML searching is LCA/ELCA method. A bottleneck for the LCA/ELCA based XML keyword search is that it is not a fast approach. In this paper, we propose a combination of ELCA approach and MapReduce technique for fast and efficient searching. A distributed LCA and ELCA computation method is designed, where each net node computes LCA and ELCA independently and only a little information needs to be transmitted.The paper employs Hadoop as the platform and implements ELCA algorithm in MapReduce framework. As a future work, new MapReduce algorithms can also be combined with other searching techniques like MLCA search techniques, Fuzzy type-ahead searching and so on.

REFERENCES

[1] Jeetendra G. Kapase and Sharmila M. Shinde, "User-Friendly Keyword Based Search on XML Data", International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.
[2] Yong Zhang, Quanlin Li and Bo Liu, "MapReduce Implementation of XML Keyword Search Algorithm", IEEE International Conference on SmartCity/SocilCom/SustainCom (2015).
[3] Y. Xu and Y. Papakonstantinou, "Efficient LCA Based Keyword Search in XML Data," Proceedings of International Conference on Extending Database Technology: Advances in Database Technology (EDBT), pp. 535-546, 2008.
[4] H. Bast and I. Weber, "Type Less, Find More: FastAuto completion Search with a Succinct Index," Proc.Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
[5] G. Li, J. Feng, and L. Zhou, "Interactive Search in XmlData," Proc. Int'l Conf. World Wide Web (WWW), pp.1063-1064, 2009
[6] L. Guo, F. Shao, C. Botev, and J. Shanmugasundaram, "Xrank: Ranked Keyword Search over Xml Documents," Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 16-27, 2003.
[7] Zhifeng Bao, Tok Wang Ling and Jiaheng Lu, "Effective XML Keyword Search with Relevance Oriented Ranking", IEEE 25th International Conference onData Engineering, 2009.
[8] V. Hristidis, N. Koudas, Y. Papakonstantinou, and D. Srivastava, "Keyword Proximity Search in Xml Trees", IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 4, pp. 525-539, April 2006.
[9] Rui Zhou, Chengfei Liu and Jianxin Li, "Fast ELCA Computation for Keyword Queries on XML Data", International Conference on Extended Database Technology (2010).
[10] Shankar Ganesh Manikandan and Siddarth Ravi, "Big Data Analysis Using Apache Hadoop", IEEE Conference on IT Convergence and Security (ICITCS), 2014.