Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Application Tools for Utility Based Web Content Mining

Rajesh Shah[1], Suresh Jain[2]
Department of CSE
Mewar University,
Chhittorgarh, Rajasthan,
India

Abstract: Extracting the useful information from the web is very important task. Web mining is the application of the data mining techniques. Web mining is defined as the extract valuable information from the web. Web mining is classified into three types: Web content mining, web structure mining and web usage mining. In this project, I used the MVC based software design patterns to implement the web content mining process. This process includes five phases. The first phase is to select the source URL, second phase is to apply tools to get the source data from the source URL,third phase is to apply techniques to clean text data, fourth phase is to display the contents to the users, fifth phase is to the part-of-speech options & sixth phase is to generate the final output.

Keywords - web mining, web content mining, web usage mining, python tools

I. INTRODUCTION

The advancement in the technology covered faster communications. The previous decade experienced a dramatic development in computer technology, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet and this technique was named as web mining. Web mining is further classified into three types: Web Content Mining, Web Structure Mining & Web Usage Mining . Using the objects like text, pictures, multimedia etc. content mining is done in the web. In web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users and the study of this navigational pattern will trace out the interest of the users [1].

Overview of Web Mining: Web mining means to discover the information from World Wide Web and it also find out its usage patterns .Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

Web mining should be decomposed into these subtasks:

1. The task of retrieving intended Web documents (Resource finding).
2. Automatically selecting and pre processing specific information from retrieved Web Resources (Information selection and pre processing).
3. Automatically discovers general patterns at individual Web sites as well as across multiple sites (Generalization).
4. Validation and/or interpretation of the mined patterns (Analysis).

Categories of Web Mining: Web mining is categorized into three areas of interest based on part of Web to mine:

1. Describes discovery of useful information from contents, data and documents (Web content mining). Two different points of view: Information Retrieval view and Data Base view.
2. Model of link structures, topology of hyperlinks Categorizing of web pages (Web structure mining).
3. Mines secondary data derived from user interactions (Web usage mining).

II. WEB CONTENT MINING

Web Content Mining is the process of extracting useful information from the content of Web documents. Logical structure, semantic content and layout are contained in semi-structured webpage text. Topic discovery, extracting association patterns, clustering of Web documents and classification of Web pages are some of research issues in text mining. These activities use techniques from other disciplines – IR (Information Retrieval), IE (Information Extraction), NLP (Natural Language Processing) and others. Automatic extraction of semantic relations and structures from Web is a growing application of web content mining. In this area, several algorithms are used. Hierarchical clustering algorithms on terms in order to create concept hierarchies, formal concept analysis and association rule mining to learn generalized conceptual relations and automatic extraction of structured data records from semi-structured HTML pages. In contrast to unstructured texts, structured data is also easier to extract. This problem has been studied by researchers in Artificial Intelligence and database and data mining [3] [4].

Web Content Mining Techniques: It identifies the useful information from the web contents/data/documents, however, such a data in its broader form has to be further narrowed down to useful information. Web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. Here, the several approaches in web content mining are represented [5].

Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

1. Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence one could consider text mining as an instance of web content mining to provide effectively exploitable results, preprocessing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques. Content mining has been accomplished on unstructured data such as text. Mining of unstructured data provides unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic content mining is a type of text mining. Some of the useful techniques used in text mining are as Information Extraction, Information Visualization, Topic Tracking, Summarization, Categorization, and Clustering (Unstructured Data Mining Techniques).
2. The techniques which have been used for mining structured data are referred as Structured Data Mining (Structured Data Mining Techniques).
3. The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language (Semi-Structured Data Mining Techniques).

Some of the Multimedia Data Mining Techniques are SKICAT, Multimedia Miner, Color Histogram Matching and Shot Boundary Detection (Multimedia Data Mining Techniques) [6].

III. WEB USAGE MINING

The web usage mining generally includes the following several steps: data collection, data pretreatment, and knowledge discovery and pattern analysis.

A] Data collection:

Web Usage Mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time. Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs[6].

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories below in the figure [6].

B] Data preprocessing:

Web Usage Mining in cloud computing is one of the categories of data mining technique that identifies usage patterns of the web data, so as to perceive and better serve the requirements of the web applications. The working of WUM involves three steps - preprocessing, pattern discovery and analysis. The first step in WUM - Preprocessing of data is an essential activity which will help to improve the quality of the data and successively the mining results. This research paper studies and presents several data preparation techniques of access stream even before the mining process can be started and these are used to improve the performance of the data preprocessing to identify the unique sessions and unique users in cloud computing . The methods proposed will help to discover meaningful pattern and relationships from the access stream of the user and these are proved to be valid and useful by various research tests. The paper is concluded by proposing the future research directions in this space [6].

In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

1) Data Cleaning:

The most important task of the Web Usage Mining in cloud computing process is data preparation. This process is diagrammatically represented in Fig . The success of the project is highly correlated to how well the data preparation task is executed. It is of utmost importance to ensure, every nuance of this task is taken care of. This process deals with logging of the data; performing accuracy check; putting the data together from disparate sources; transforming the data into a session file; and finally structuring the data as per the input requirements. The data used for this project is from the RIT Apache server logs, which is in the Common Log File format. This access log includes the agent and the referrer in the data as one of the attributes.

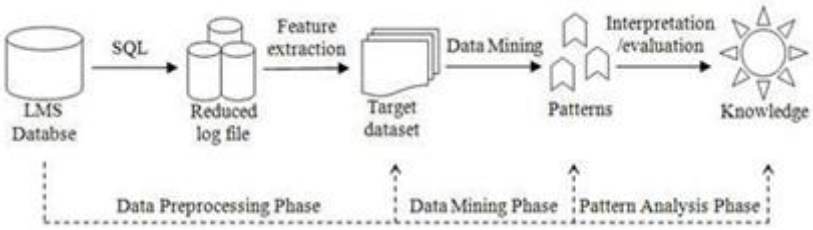


Fig: data preparation

2) Path Completion

An implementation of data preprocessing system for Web usage mining and the details of algorithm for path completion are

presented. After user session identification, the missing pages in user access paths are appended by using the referrer-based method

which is an effective solution to the problems by using proxy servers and local caching. The reference length of pages in complete

path is modified by considering the average reference length of auxiliary pages which is estimated in advance the maximal forward

references and the reference length algorithms. As verified by practical Web access log, the proposed path completion algorithm

efficiently appends the lost information and improves the reliability of access data for further Web usage mining calculations[6].

C] Knowledge Discovery:

In general, knowledge discovery can be defined as the process of identifying interesting new patterns in data.

These patterns can be e.g., relations, events or trends, and they can reveal both regularities and exceptions[6].

D] Pattern analysis:

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user.

First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry

on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic

service to the electronic commerce website [6].

Web Mining Process

1. It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web (Information Retrieval).
2. It is the task of automatically selecting and pre-processing specific information from retrieved web resources (Pre-processing).
3. It is the task to automatically discover general patterns of individual web sites as well as across multiple web sites (Pattern Recognition & Machine Learning).
4. It is the task of analyzing, validating and interpreting the mined patterns (Analysis) [7].

IV. PROBLEM STATEMENT

Users could encounter following problems when interacting with the Web.

a) Most people use some search service when they want to find specific information on the Web. A user usually inputs a simple keyword query and a result is a list of ranked pages. This ranking is based on their similarity to the query. Today's search tools have some problems are Low precision and low recall, mainly because of wrong or incomplete keyword query. This leads to irrelevance of many search results (Finding relevant information).

b) This problem is data-triggered process that presumes that we have a collection of web data and we want to extract potentially useful knowledge from these data (Creating new knowledge).

c) People differ in the contents and presentations they prefer while interacting with the Web (Personalization of information).

d) This is a group of sub-problems such as mass customizing information to intended consumers, problems related to effective web site design and management, problems related to marketing and others (Learning about consumers or individual users) [8].

This method focuses on the two objectives –

The objectives are -

- Focusing on the architecture design for the efficient web mining.
- Finding the required patterns in the source content.

Work Process

The following steps are:

- Select the source URL
- Apply tools to get the source data from the source URL
- Apply techniques to clean text data
- Display the contents to the users
- Select the part-of-speech options
- Generate the output file

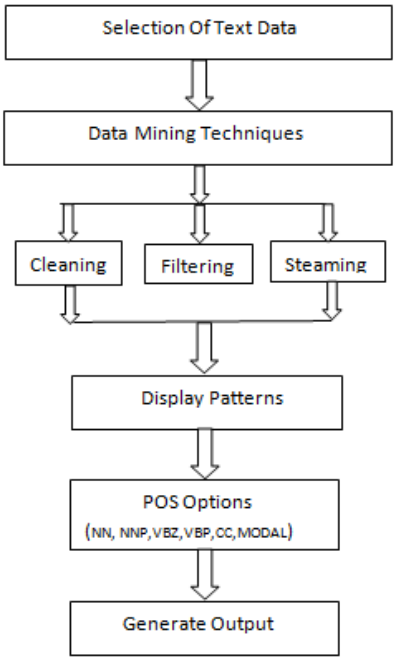
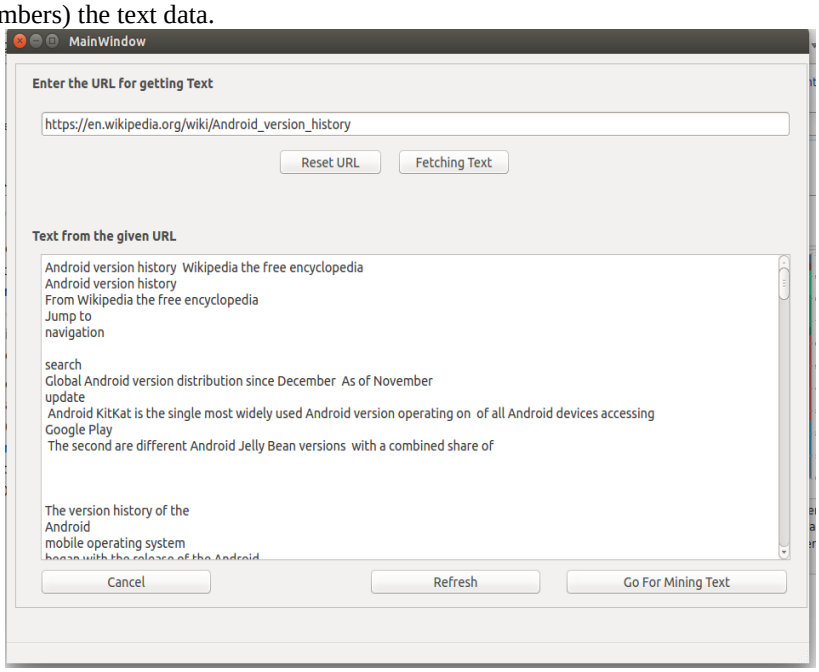


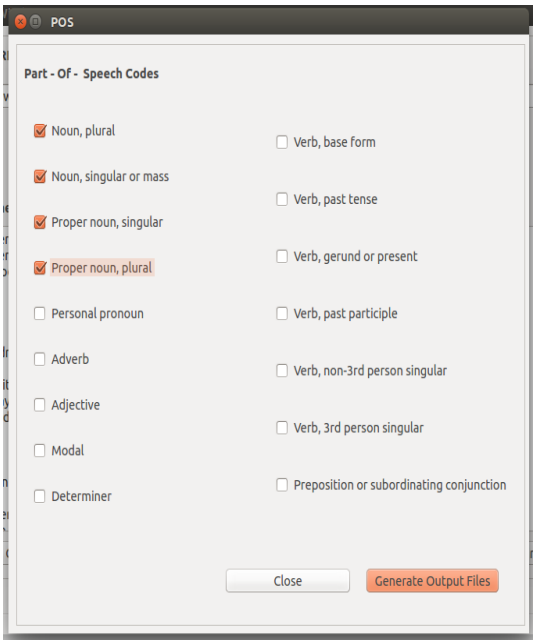
Fig: Web Mining Process

V. EXPERIMENTAL RESULTS & DISCUSSION

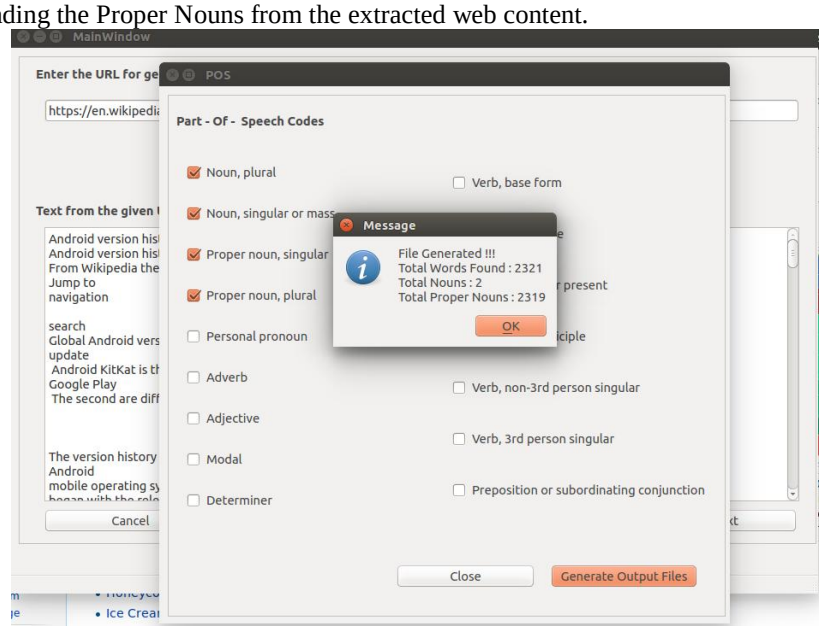
To implement this process, I have used PyQt (A Cross-Platform Application Development Framework), Python as a developing language, NLTK for text processing, BeautifulSoup for extract information. The working process is shown in below screenshots. First we select the source URL to extract the content from that URL. Now the click on "Fetching Text" button, our code will open that URL and extract the html part of that web page & apply the data mining techniques like – cleaning, filtering & steaming on that html part. These techniques are to clean (JavaScript, CSS Style sheet, Special characters, Numbers) the text data.



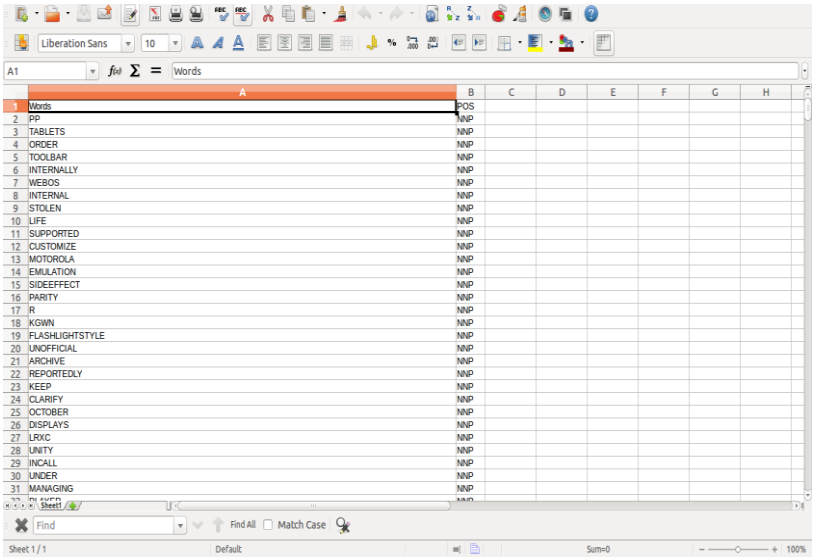
Now when we click on "Go For Mining Text" button, a dialog box is open which contain all the part- of- speech code options. We select the any part-of-speech code options by clicking the check box in front of them it will extract only that data from the entire text. And finally the final data will save into a file by clicking on "Generate Output Files".

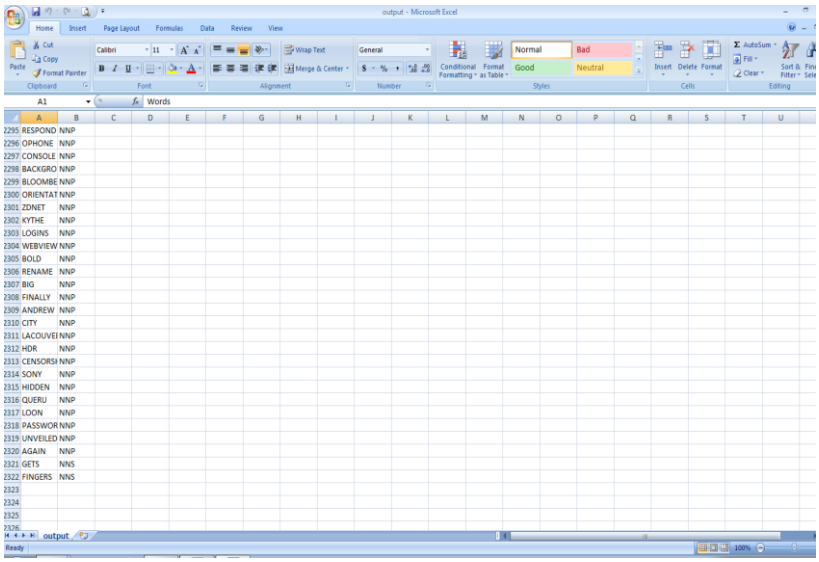


In the output file, total words found is 2321, in which 2319 words are proper nouns & 2 words are nouns. I have achieved 100% efficiency in finding the Proper Nouns from the extracted web content.



Generated Output File –





VI. RELATED WORKS

Various scholars and researches have proposed related work in web content mining, which are discussed below -

Aidan Finn discusses in his research work Fact or fiction, Content classification for digital libraries, methods for content extraction from single-article sources, where content is supposed to be in a single body. The algorithm tokenizes a page into either words or tags. The page is sectioned into three contiguous regions, placing boundaries to partition the document such that most tags are placed into outside regions and word tokens into the center region. This approach works well for single-body documents, but destroys the structure of the HTML and doesn't produce good results for multi-body documents, i.e., where content is segmented into multiple smaller pieces like we find on Web Blogs [9].

McKeon in the NLP (Natural Language Processing) group at Columbia University detects the largest body of text on a webpage (by counting the number of words) and classifies that as content. This method works well with simple pages. However, this algorithm produces noisy or inaccurate results handling multi-body documents, especially with random advertisement and image placement.

Arvind kumar Sharma & P.C. Gupta, the main uses of web content mining are together, categorize ,organize & provide the best possible information available on the www to the user requesting. Future scope of web content mining includes predicting users needs in order to improve the usability, scalability and user retention [10].

Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi the Web Data Mining Tools are primordial to scanning the many HTML documents, images, and text provided on Web pages. The result is provided to the search engines, in order of relevance giving more productive results of each search. In this paper, we presented a non-exhaustive list of the available Web Content Mining Tools. Through this study, we established some objective criteria for comparison. Based on these criteria, we gave a comparative table of these different tools. We believe that research in Web mining is promising as well as challenging, and this field will help produce applications that can more effectively and efficiently utilize the Web of knowledge. We are currently working to design and implement a Web mining system based on multi-agents technology. We pretend that such system reduce the information overload and search depth. That is helpful to users using the web within a platform for ecommerce or eLearning [11].

VII. CONCLUSION

We have applied mining techniques to unstructured text is termed as text mining. Text mining is an instance of web content mining. I have done successfully scraping on the web content by using python language. The results after the analysis contained valuable information from the web content depending on the selected point of speech (POS) options. I found this technique is efficient & easy to generate the outputs.

REFERENCES

[1] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.

[2] D. Jayalatchumy, Dr. P.Thambidurai Web Mining Research Issues and Future Directions – A Survey IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3 (Sep. - Oct. 2013), PP 20-27 www.iosrjournals.org

[3] G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.

[4] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. Tapping into the Power of Text Mining. Communications of the ACM - Privacy and Security in highly dynamic systems. Vol. 49, Issue-9.

[5] Gupta,V. and Lehal,G.S. 2009.A Survey of Text Mining Techniques And Applications.Journal Of Emerging Technologies In Web Intelligence.Vol. 1,pp.60-76.

[6] Rajesh shah and Suresh Jain, Web Mining Using Cloud Computing Technology, International Journal of Scientific Research in Computer Science and Engineering, Volume- 3, Issue-2 ISSN: 2320-7639.

[7] Pol, K., Patil, N., Patankar, S. and Das, C. 2008. A Survey on Web Content Mining and extraction of Structured and Semi structured Data. IEEE First International Conference on Emerging.

[8] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.

[9] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648. Morgan Kaufmann Publishers, 2006.

[10] A. F. R. Rahman, H. Alam and R. Hartono. Understanding the Flow of Content in Summarizing HTML Documents. In Int. Workshop on Document Layout Interpretation and its Applications, DLIA01, Sep., 2001.

[11] Arvind Kumar Sharma and p.c. gupta,Study & Analysis Of Web Content Mining Tools To Improve Techniques Of Web Data Mining, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012.

[12] Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi, Overview Of Web Content Mining Tools, The International Journal Of Engineering And Science (IJES) ||Volume||2 ||Issue|| 6 ||Pages|| 106-110||2013|| ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805.

[13] Rahesh Shah et al , "Web Mining Using Cloud Computing Technology", "International Journal of Scientific Research in Computer Science and Engineering", Volume-3, Issue-2 ISSN: 2320-7639.

[14] Makhan Kumbhkar et al ," Analysis of Cloud Computing Security Issues in Software as a Service", "International Journal of Scientific Research in Computer Science and Engineering", Volume-2, Issue-3 ISSN: 2320-7639.

[15] Rahesh Shah et al ," A Comparative Study of Two Software Development Approaches: Traditional and Object Oriented "," International Journal of Advanced Research in Computer Science and Software Engineering" Volume 5, Issue 5, May 2015.

[16] Makhan Kumbhkar et al ," Security in Cloud Environment", ", "International Journal of Scientific Research in Computer Science and Engineering", Volume-2, Issue-3 ISSN: 2320-7639.

[17] Makhan Kumbhkar et al ," Security Analyzing in UNIX for Cloud Computing Environment"," International Journal of Innovative Research in Computer and Communication Engineering", Vol. 3, Issue 10, October 2015.

