International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Review of Web of Things and Request Dependency Graph Analysis and Mining

Preeti Gungrao Patil[1], Mansi Bhonsle[2]

[1] PG Scholar, Department of Computer Engineering, G.H.Raisoni College of Engg. and Mgmt., Pune, India-412 207

[2]Assistant Professor, Department of Computer Engineering, G.H. Raisoni College of Engg. and Mgmt., Pune, India-412 207

Abstract: ***Web of things is an application layer of internet of things where real objects communicate with each other when they get connected with World Wide Web (www). It describes approaches and programming patterns to support internet of things. Here, it is described how the web objects (links or web pages) of a website talk to each other, what is the flow between them, pattern formed when they get linked up and how users use it. This helps in analysis of the website and determines the scope of improvement to make it more efficient and useful.***

Keywords: Internet of Things, Web log mining, Request Dependency Graph, Primary Request.

1. Introduction

'Internet', one word which is buzzing around the world and which is reaching to its new level day by day changing the way world operates! With introduction of World Wide Web concept, there has been dynamic movement in each field of business. Reaching people in any part of the world is at finger tips, creating a platform for knowledge sharing and growth of society. Now a new era of internet is arriving with the preamble of Internet of Things where objects would start communicating with each other using internet connectivity. [10] Here objects are embedded with software, sensors, electronics and internet connectivity to gather and swap information. Still there should be something which would allow these objects to connect to World Wide Web. This is done with the help of concept known as Web of Things. There would be a great boost in the field of monitoring and research due to this technological shift.

The Web of Things (WoT) is a computing concept that describes a scope where everyday things are fully integrated with the Web. [11] The prerequisite for WoT is for the "things" to have embedded computer machines that enable interaction with the Web. Such smart devices would then be able to communicate with each other using present Web standards.

Considered a subset of the Internet of Things (IoT), WoT focuses on software standards and frameworks such as REST, HTTP and URIs to generate applications and services that combine and interact with a variety of network devices. So, you could think of the Web of Things as day-today objects being able to access Web services. The key point is that this doesn't involve the reinvention of the modes of communication because existing standards are utilized.

Communications
Connected
Managed
Secured
Medical
Smart homes
Industrial
Smart Grid
Building Automation
Vehicles

Figure 1: Web of Things concept

Internet of Things is more often used in the context of radiofrequency identification (RFID) and how physical objects are attached to the Internet and can communicate with each other. Both terms are difficult to define precisely, although they are related in their general pattern. [1] In this paper we do a survey of papers on how the information is retrieved from web of things and how it is being utilized and analyzed to improve the efficiency and performance of web-sites containing different web objects (web-pages).

2. Basic Concepts

2.1 Web Mining

Data mining is the practice of monitoring large pre-existing databases in order to generate new information. Web mining is the application of **data mining** techniques to identify patterns from the World Wide Web (www). World Wide Web is a huge storehouse of web pages and links. It offers large quantity of data for the Internet users. The growth of web is phenomenal as around one million pages are added per day. Users' accesses are recorded in web logs. Web usage mining is a part of mining techniques in logs. [8] In the same manner web content mining, web usage mining and structure mining is carried out to understand the structure and content pattern of web pages.

2.2 Request Dependency Graph

In Web of Things environment, web traffic logs contain valuable information of how people interact with smart devices and web servers. Mining the wealth of information available in the web access logs has theoretical and practical significance. Mining web logs involves modeling the relationships among HTTP requests. [3] It provides a graphical representation of behavior of web clients.

Web logs at server side are analyzed on the basis of statistical and structural properties of complete web environment. More challenges faced when web data grows in volume. Analysis on variety of web logs becomes tedious. A system is developed a system where analysis on the web log will be applied and productivity can be increased. Efficient algorithm (Request Dependency Graph) used to model the relationships among HTTP requests. Generate a graph by mining the temporal and causal information among aggregated HTTP requests. Design and implementation of an algorithm for primary requests identification is carried out.

2.3 Primary HTTP Request Identification

The set of initial HTTP requests of web pages is only a small part of the total web traffic. Primary requests identification is a process of obtaining the initial set of requests triggered by users or devices from a large number of captured HTTP requests. It is to identify the initial HTTP requests of opening a web page, which is triggered by a user click or a device access action, from the captured web traffic logs. [4] We use the request dependency graph model to describe the complex web browsing behavior. Based on the graph, we propose a two-step algorithm to identify the primary requests from a huge number of HTTP requests of web pages and embedded objects.

3. Sources of Data

To apply request dependency graph algorithm and perform analysis we need data in first place. The major source for obtaining data is websites, software, facts, figures etc. Huge data can be found about company, consumers, producers, retailers, legal documents, data warehouse etc. Data sources are given below.

3.1 Web Servers

The Server Side Web servers are without a doubt the wealthiest and the most widely recognized wellspring of information. They can gather a lot of data in their log records and in the log documents of the databases they utilize. These logs for the most part contain essential data e.g. name and IP of the remote host, date and time of the solicitation, the solicitation line precisely as it originated from the customer, and so forth. This data is typically spoken to in standard organization e.g.: Common Log Format, Extended Log Format, and LogML. At the point when abusing log data from web servers, the significant issue is the recognizable proof of clients' sessions. Aside from web logs, clients' conduct can likewise be found on the server side by method for TCP/IP bundle sniffers.

3.2 Proxy Servers

Many network access suppliers (ISPs) provide for their client Proxy Server administrations to enhance route speed through storing. In numerous regards, gathering route information at the intermediary level is essentially the same as gathering information at the server level. The principle contrast for this situation is that intermediary servers gather information of gatherings of clients getting to immense gatherings of web servers.

3.3 Web Clients

The Client Side usage information can be followed additionally on the customer side by utilizing Java Script, java applets, or even adjusted programs. These methods maintain a strategic distance from the issues of clients' session caching so as to distinguish proof and the issues brought about (like the utilization of the back catch). What's more, they give itemized data about real client practices. In any case, these methodologies depend vigorously on the clients' collaboration and rise numerous issues concerning the security laws, which are entirely strict.

4. Web Mining Methodologies

It is the process of extracting useful resources or information. Data mining is analyzing data from different perspectives and summarizing it into useful information. Mined information can be used to increase revenue, cuts costs, or both. Web Mining is the application of data mining techniques that discovers patterns from the World Wide Web. It is focused on learning about web user with their interaction with web sites. It is used to extract knowledge from WWW and provides value to data from huge volume of web logs. Web Mining can be categorized into following techniques.

4.1 Web Usage Mining

Discover interesting usage patterns from web usage data, 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. Obtains correlations between the various pages visited during a browsing session. [8] Web usage mining is performed on Web Server Data, Application Server Data and Application Level Data. Data mined from Web utilization information permits rebuilding and better administration of the webpage, giving more viability to it. Following are the various phases of web usage mining.

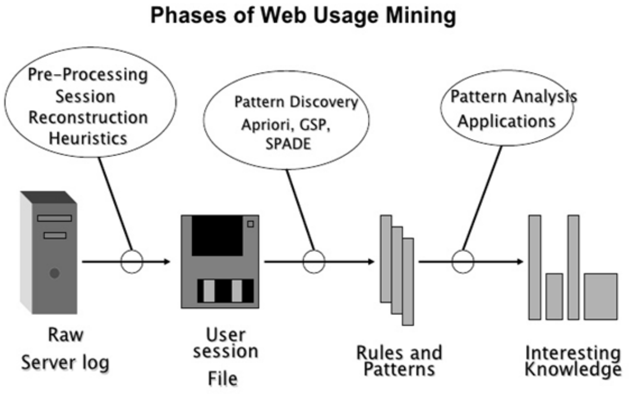

Phases of Web Usage Mining
Pre-Processing Session Reconstruction Heuristics
Pattern Discovery Apriori, GSP, SPADE
Pattern Analysis Applications
Raw Server log
User session File
Rules and Patterns
Interesting Knowledge

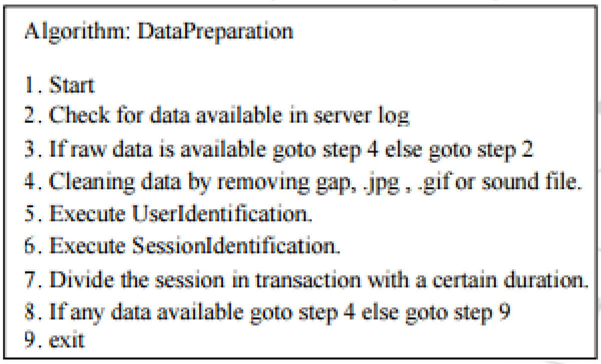
Figure 2: Phases of Web Usage Mining

4.1.1 Data Preparation

In Data Preparation phase the web log data must be cleaned, filtered, integrated and transformed in such a way that the irrelevant and redundant data can be removed, user session and transaction can identified.

a) Data cleaning

The first step of data preparation is data cleaning or filtering. It is very important as there have many unnecessary entries in the log files. Elimination of irrelevant items can be accomplished by checking the suffix of the URL name, which tells one what format these kinds of files are. For example, the surrounded graphics can be filtered out from the Web log file, whose suffix is usually the form of “gif”, “jpeg”, “jpg”, “GIF”, “JPEG”, “JPG”, can be removed. In the same way the unwanted sound files can be removed.



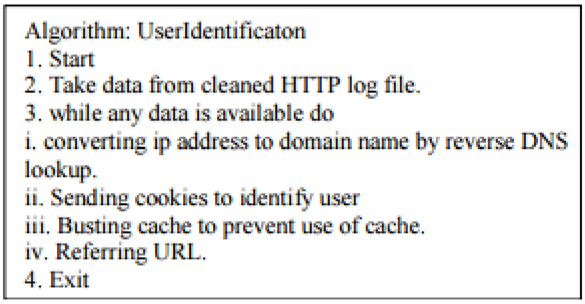
```
Algorithm: DataPreparation

1. Start
2. Check for data available in server log
3. If raw data is available goto step 4 else goto step 2
4. Cleaning data by removing gap, .jpg , .gif or sound file.
5. Execute UserIdentification.
6. Execute SessionIdentification.
7. Divide the session in transaction with a certain duration.
8. If any data available goto step 4 else goto step 9
9. exit
```

b) User Identification

Once HTTP log files have been cleaned, next step in the data preparation is the identification of the user, through heuristics.

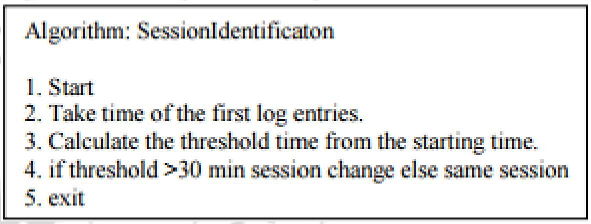
1) By converting ip address to domain name exposed some knowledge. For example, one can estimate where visitors live by looking at the extension of each visitor’s domain name, such as .ca (Canada); .au (Australia); cn (China), etc.
2) The web server randomly assigned an Id to the web browser while it connects first time to the site. This is called cookies. The Web browser sends the same ID back to the Web server, effectively telling the Web site that a specific user has returned. Cookies help the Web site developer to easily identifying individual visitors, which results in a greater understanding of how the site is used. Cookies also help visitors by allowing Web sites to recognize repeat visits.
3) Cache prevents much user access to be recorded in the log file when a page hit by the user already in the cache. Cache busting is one solution of this problem.



```
Algorithm: UserIdentificaton
1. Start
2. Take data from cleaned HTTP log file.
3. while any data is available do
    i. converting ip address to domain name by reverse DNS lookup.
    ii. Sending cookies to identify user
    iii. Busting cache to prevent use of cache.
    iv. Referring URL.
4. Exit
```

c) Session Identification

Session identification can be performed using time interval between consecutive log entries. If two accesses from the same user are separated by an interval longer than a threshold they considered as different session. Sometimes threshold considered as 30 minutes time interval. Another way to identify session is using a time out to identify the end of the session. After data preparation the server log file data have to be prepared for pattern discovery. This data is more organized, classified which we called web warehousing.



```
Algorithm: SessionIdentificaton

1. Start
2. Take time of the first log entries.
3. Calculate the threshold time from the starting time.
4. if threshold >30 min session change else same session
5. exit
```

4.1.2 Pattern Discovery

After data preparation phase, the pattern discovery method should be applied. This phase consists of different techniques derived from various fields such as statistics, machine learning, data mining, pattern recognition, etc. applied to the Web domain and to the available data. The task for discovering the patterns offer some techniques as statistical analysis, association rules, sequential pattern analysis, clustering and so on. Here we will briefly describe some techniques to discover patterns from processed data. To determine the visitor’s location converting the IP address into its domain name is a good way. Looking up the extension of the domain name one may figure out the country of the visitor.

4.1.3 Pattern Analysis

Involves analysis of patterns discovered in above steps to judge their interesting nature. These log data can be used in web site designing, modifying and also to improve the overall performance of web site. Pattern analysis plays a crucial role in predicting the business output in numbers. All the post actions to improve the online application depend on graphs and flowcharts generated by pattern analysis step.

4.2 Web Content Mining

It is process of information or resource discovery from content of millions of sources across the World Wide Web. Web content consists of text, image, audio, video, metadata and hyperlinks. The heterogeneity and the lack of structure of expanding web data is a challenge for data retrieval from web pages. Scanning and mining of text, graphs and pictures from a Web page is done. This find out the significance of content to the search query. [6]

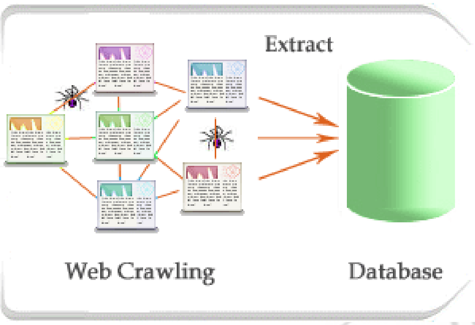


Figure 3: Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. It describes the discovery of useful information from the web documents. In web content mining the content may be text, image, audio, video, metadata and hyperlinks etc. Web content mining also distinguishes personal home pages with other web pages. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages.

4.3 Web Structure Mining

It uses the hyperlink structure of the Web as an (additional) information source. It generates the structural summary for the web sites and web pages. The aim of the web structure mining is to generate the structural abstract about the websites and webpage. It establish the link construction of the hyperlinks at the inter text level. The topology used in web structure mining is that will categorize the web pages and spawn the information like similarity and relationship between the different websites.

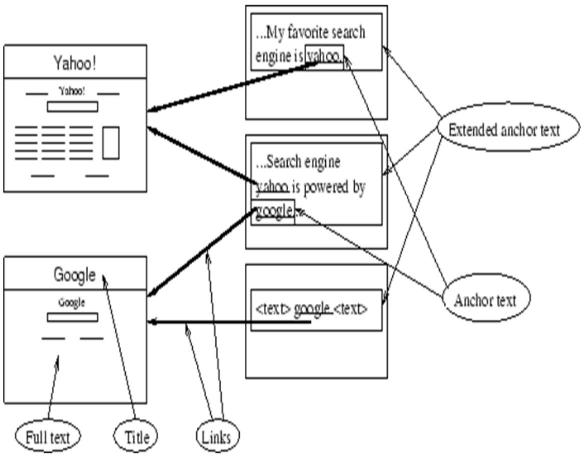


Figure 4: Web Structure Mining

Web structure mining is the process of analyzing the hyperlink and mine important information from it and steps to achieve the information is tedious one. Likewise the remaining also used to mine the structure of document, analyze the structure of page and to describe the HTML format or XML usage. The primary objective of the Web Structure Mining is to generate the structural synopsis about the Web site and Web page. Web Structure mining will sort out the Web pages in different category and from the category to generate the information like the similarity and relationship between different Web sites. The type of mining can be either performed at document level called as Intra-page and at the same time another level is performed at hyperlink level called inter-page mining. The challenge for web structure mining is to deal with the structure of the hyperlinks within the web itself.

5. Types of Graphs

Graph is a requested pair G = (V, E) including a set V of vertices, hubs or focuses together with a set E of edges, circular segments or lines, which are 2-component subsets of V (i.e., an edge is connected with two vertices, and the connection is spoken to as an unordered pair of the vertices as for the specific edge). To maintain a strategic distance from equivocalness, this sort of diagram may be portrayed exactly as undirected and straightforward. Different types of graphs are explained below.

5.1 Undirected graph

An undirected diagram is a chart in which edges have no introduction. The edge (x, y) is indistinguishable to the edge (y, x), i.e., they are not requested sets, but rather sets {x, y} (or 2-multisets) of vertices. The most extreme number of edges in an undirected diagram without a circle is $n(n - 1)/2$.

5.2 Coordinated graph

A coordinated diagram or digraph is a chart in which edges have introductions. It is composed as a requested pair G = (V, An) (occasionally G = (V, E)) with

V a set whose components are called vertices, hubs, or focuses;

An arrangement of requested sets of vertices, called bolts, coordinated edges (some of the time basically edges with the relating set named E rather than A), coordinated bends, or coordinated lines.

A bolt (x, y) is thought to be guided from x to y; y is known as the head and x is known as the tail of the bolt; y is said to be an immediate successor of x and x is said to be an immediate antecedent of y. In the event that a way leads from x to y, then y is said to be a successor of x and reachable from x, and x is said to be an ancestor of y. The bolt (y, x) is known as the reversed bolt of (x, y).

5.3 Blended graph

A blended diagram is a chart in which a few edges may be coordinated and some may be undirected. It is composed as a requested triple G = (V, E, A) with V, E, and A characterized as above. Coordinated and undirected charts are uncommon cases.

5.4 Bipartite graph

A bipartite diagram is a chart in which the vertex set can be parceled into two sets, W and X, so that no two vertices in W share a typical edge and no two vertices in X share a typical edge. On the other hand, it is a chart with a chromatic number of 2.

In a complete bipartite diagram, the vertex set is the union of two disjoint sets, W and X, so that each vertex in W is contiguous each vertex in X yet there are no edges inside of W or X.

Request Dependency Graph

6.1 Example of request dependency graph

The Web of Things environment, web activity logs contain significant data of how individuals communicate with keen gadgets and web servers. Mining the abundance of data accessible in the web access logs has hypothetical and commonsense hugeness for some vital applications like system improvement and security administration. The main basic stride of the mining errand is displaying the connections among HTTP asks for getting to web articles to research the conduct of web customers. In this paper, we present the solicitation reliance diagram, a chart representation of the connections among HTTP asks. Theoretically, a coordinated connection from A to B in the chart implies that the getting to of web item B is created by the getting to of An, i.e., B relies on upon A. We propose an approach to build up such a diagram by mining the fleeting and causal data among accumulated HTTP asks.

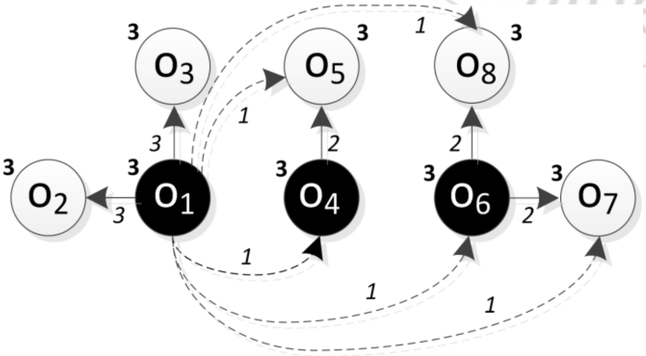


Figure 5: Example of request dependency graph

Based on the web browsing process and basic concepts described above, we introduce a dependency graph model to depict the dynamic web browsing behavior. Formally, we model the browsing behavior as a directed and weighted graph, G = (O,S,E,W), referred to as a Request Dependency Graph (RDG).

O = {o1,o2,….on} is the set of nodes representing accessed objects which are identified by URLs.

Each node oi is assigned an occurrence count S[oi] € S of the accessed object oi.

E is the set of directed edges with weights W. There is an edge from node oi to oj if and only if they meet the following conditions:

i. For a request sequence v= {ri-1,ri,….,rj} generated by the same device, in which accessed objects are {oi-1,oi,….,oj}, the interval between the accessing time of ri-1 and ri is larger than Ŧ, where Ŧ is called the lookahead time window.
ii. In the sub-sequence v'= {ri,….,rj} of v, the interval between each pair of adjacent requests is smaller than Ŧ.

A directed edge from oi to oj represents the dependency relationship between them. The weight of a directed edge is the number of times the pair < oi; oj > appears in the measured HTTP request sequence R. A directed edge and the weight of the edge indicate the access of oj followed by the access of oi and the number of occurrences of such an action, respectively. Ideally, if all devices open web pages one by one with a given stay time and Ŧ equals to the stay time, the edge from oi to oj indicates that oj is an embedded object in the web page oi. the request dependency graph derived from the web traffic log is more complicated than the simple structure of objects relationships defined by developers of the accessed web servers.

6.2 Web log processing

Web Usage Mining addresses the issue of separating behavioral examples from one or more web access logs [3]. the whole process can be isolated into three noteworthy steps. The initial step, pre-preparing, is the errand of precisely recognizing pages got to by web guests. This is an exceptionally troublesome errand in view of page reserving and gets to by web crawlers. The second step, design revelation, includes uses of information mining calculations to the pre-prepared information to find designs. The last step, design investigation, includes examination of examples found to judge their interestingness.

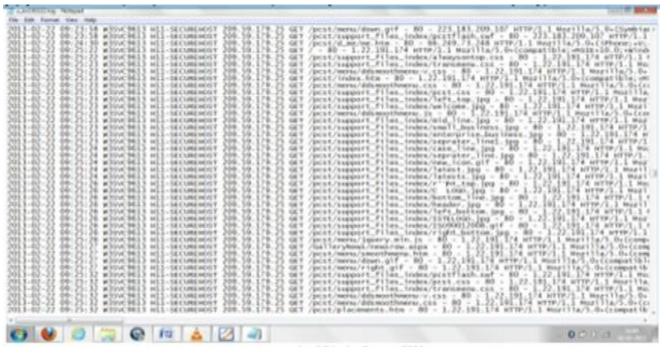


Figure 6: Web log file

Web server records every one of clients' exercises of the site as web servers Logs. Most log records have content arrangement and every log section is spared as a line of content. There are numerous sorts of web logs, for example, NCSA position, W3C design and IIS group, yet they have the same fundamental data. Log information spoke to in W3C amplified configuration is appeared in figure1. These log information can be utilized as a part of site planning, changing furthermore to enhance the general execution of site. Subsequent to recognizing the diverse web server log information documents there is a need to blend the log records.

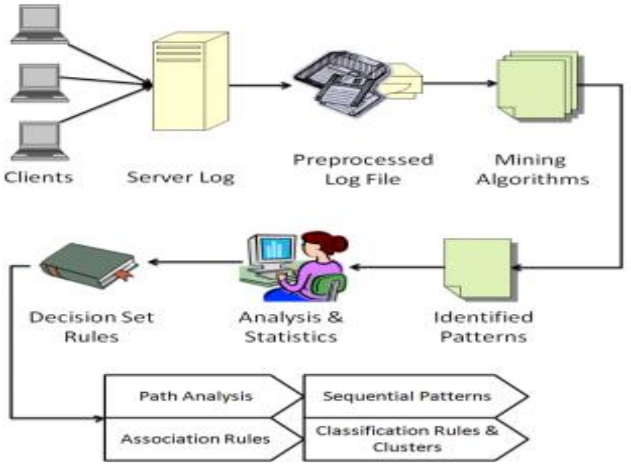


Figure 7: Web log processing

6. Applications

Excitement about the web in the past few years has led to the web applications being developed at a much faster rate in the industry than research in web related technologies.

A host of web mining techniques, such as associations between pages visited and click-path analysis are used to improve the customer’s experience during a “store visit.” Knowledge gained from web mining is the key intelligence behind Amazon’s features such as “instant recommendations,” “purchase circles,” “wish-lists,” etc. Google is one of the most popular and widely used search engines. It provides users access to information from over 2 billion web pages that it has indexed on its server. The quality and quickness of the search facility makes it the most successful search engine. Earlier search engines concentrated on web content alone to return the relevant pages to a query. [7] Google was the first to introduce the importance of the link structure in mining information from the web. PageRank, derived from RDG which measures the importance of a page, is the underlying technology in all Google search products, and uses structural information of the web graph to return high quality results.

“Web-wide tracking,” i.e. tracking a visitor across all web pages he visits with the help of unique identification number, is an intriguing and controversial technology. [9] It can provide an understanding of an individual’s lifestyle and habits to a level that is unprecedented, which is clearly of tremendous interest to marketers. Primary request identification plays a crucial role in identifying a new user. In this manner RDG along with primary request identification works in a web of things environment to extract useful information about the interaction of web objects over www and analysis. This initiates consistent improvement in web page design and web of things implementation.

7. Conclusion

We have used request dependency graph as base for web log mining and derived possible analysis from the result of algorithm implementation. Experimental results have substantiated that our method achieves higher accuracy as compared with the widely used data cleaning method. Browsing behavior modeling and primary requests identification are fundamentally critical for subsequent web usage mining. RDG along with primary request identification works in a web of things environment to extract useful information about the interaction of web objects over www and analysis helping in improving the performance and efficiency of web pages in web of things.

References

[1] Jun Liu, Member, IEEE, Cheng Fang, and Nirwan Ansari, Fellow, IEEE “Request Dependency Graph: A Model for Web Usage Mining in Large-scale Web of Things” DOI 10.1109/JIOT.2015.2452964, IEEE Internet of Things Journal IEEE INTERNET OF THINGS JOURNAL, VOL. XX, NO. XX, XX 2015

[2] P. Gill, M. Arlitt, N. Carlsson, A. Mahanti, and C. Williamson, “Characterizing organizational use of web-based services: Methodology, challenges, observations, and insights,” ACM Transactions on the Web, vol. 5, no. 4, pp. 19, 2011.

[3] F. Schneider, B. Ager, G. Maier, A. Feldmann, and S. Uhlig, “Pitfalls in HTTP traffic measurements and analysis,” In Passive and Active Measurement, Springer Berlin Heidelberg, pp. 242-251, Jan. 2012.

[4] M. R. Meiss, F. Menczer, and A. Vespignani, “Structural analysis of behavioral networks from the Internet,” Journal of Physics A: Mathematical and Theoretical, vol. 41, no. 22, 2008.

[5] K. Ashton, “That ‘internet of things’ thing,” RFiD Journal, vol. 22, no. 7, pp. 97-114, 2009. [2] D. Guinard, “A Web of things application architecture,” Dissertation, Eidgenossische Technische Hochschule ETH Zurich, Nr. 19891, 2011.

[6] Frieder Ganz, Daniel Puschmann, Payam Barnaghi, Senior Member, IEEE, and Francois Carrez “A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things” IEEE INTERNET OF THINGS JOURNAL, VOL. 2, NO. 4, AUGUST 2015,

[7] Benedikt Ostermaier, Kay R¨omery, Friedemann Mattern, Michael Fahrmairz and Wolfgang Kellerer, “A Real-Time Search Engine for the Web of Things”, DOCOMO Euro-Labs, Munich, Germany

[8] R.Khanchana and M. Punithavalli, “Web Usage Mining for Predicting Users’ Browsing Behaviors by using FPCM Clustering” IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011

[9] Sayalee Narkhede and Tripti Baraskar, "HMR LOG ANALYZER: ANALYZE WEB APPLICATION LOGS OVER HADOOP MAPREDUCE", International Journal of UbiComp (IJU), Vol.4, No.3, July 2013

[10] Luigi Atzori*, Antonio Iera**, Giacomo Morabito***, and Michele Nitti: The Social Internet of Things (SIoT) - When Social Networks meet the Internet of Things:Concept: Architecture and Network CharacterizationI: Paper submitted and published in Computer Networks, Volume 56, Issue 16, 14 November 2012, Pages 3594–3608.

[11] Wikipedia, "Web of Things", https://en.wikipedia.org/wiki/Web_of_Things

[12] Wikipedia, "Internet of Things", https://en.wikipedia.org/wiki/Internet_of_Things

[13] Wikipedia, "Request Dependency Graph",https://en.wikipedia.org/wiki/Dependency_graph

[14] NASA official website https://www.nasa.gov/

Author Profile

Preeti Patil received a Bachelor of Engineering degree in the field of Computer Science from Kolhapur University in the year 2012. She worked as an Asst. Professor at Bharti Vidyapeeth College of Engineering, Kolhapur for two years. To complete her learning aspirations she decided to pursue Masters in Computer Networks from Raisoni College of Engineering and Management. Currently she is into Final semester of her ME in Computer Networks.