International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 11, November 2015 **ISSN 2319 - 4847**

Geolocation Based Web Usage Mining For Web Personalization In E-Commerce

[1]Fiona Joseph Arreppuzhekara, [2]Dr. Tanuja K. Sarode

[1] Student M.E., Thadomal Shahani Engineering College, Affiliated to Mumbai University

[2] Associate Professor, Thadomal Shahani Engineering College, Affiliated to Mumbai University

ABSTRACT

The rapid development of internet has led to businesses to leap online in order to increase their visibility & revenue by reaching out to such budding customer population which could not be possible by traditional methods. Massive amount of data are generated at the server as millions of customers interact online. To discover knowledge from this kind of data Web usage mining technique is applied. This paper proposes a geolocation based web usage mining approach to extract customer behavioral patterns from the web server logs. Customer browsing history is recorded & stored in the server log which is available only to the webmaster. From the many attributes of each web server log entry, the essential ones are extracted that will generate location based transaction details. Finally this transaction data will be fed to mining algorithm like Apriori algorithm to find association rules like what kind of products/items are majorly purchased from which geographic location, thus reflecting location based purchase pattern of customers.

Keywords:- Web usage mining, server log, geolocation, Web personalization

1. INTRODUCTION

For e-business, its sales aspect is considered as electronic commerce. E-commerce is a strategy to enter the market of business even if the company does not necessarily have a physical existence [1]. Since the advent of internet, businesses have come online and sales and purchase have become easily available to customers from the comfortable space of their homes and offices, hence customer pool has increased phenomenally. As internet is widespread many businesses have sprout online even without having a physical presence. In such a tough competitive environment it is essential to keep the customer happy to retain them.

Customers browse through web pages and this browsing history is saved in the web server log, from which potential customer interest related information can be extracted using a technology called data mining. Data mining techniques when applied on the enormous data of the World Wide Web helps to discover useful patterns which is nothing but called as web mining [2]. Web mining helps in pulling out desirable and prospective patterns and it is also used to take out information that's hidden in the web documents and from web activities [3]. As per target analysis, web mining can be classified into three different categories, which are- Web structure mining, Web content mining and Web usage mining [2].

Web structure mining applies graph theory to evaluate the nodes of the web site and also its graph structure. It is a method of determining organization data from the Web. The two levels at which web structure mining can be performed are at the document level, i.e. intra-page or at the hyperlink level, i.e. inter-page.

Web content mining as the name goes suggests mining of useful data from the web page. It also integrates the knowledge acquired from the web page content. The data displayed on the web page indicate the facts displayed for target users. This data can have structured records like table or plain text, it can have audio or images or video or even lists.

Web usage mining is the process of extracting useful information from the web server logs. It is the process of finding out what users are looking for on the Internet, by applying appropriate web mining [2].

In web usage mining the web server collects the server logs. The server maintains page request history in this server log. The server log is created automatically by the server. The latest and new page requests get added to the end of the server log as an entry. This entry based on page request consists of the customers' IP address, timestamp of the request, page or resource requested, HTTP code for the requested page whether error or success, number of bytes transferred, user agent, and referrer of the currently accessed page.

The administrative person of the website or the webmaster can only access the server logs and general internet users cannot access these files [4]. The admin uses it for the reason of perceiving "how much traffic the website is getting, how many requests have failed", and also to document and map out the customers' on-line behaviours. For instance, after performing some basic analysis on the website traffic, the server log files can help us to get answers for queries like as What pages on the website are the most hit and least popular?" [5].

To identify the actual geographical location of an object is known as Geolocation. Thus geolocation can mean that IP address of the device is mapped to a geographical location. By correlating a geographic location with the Internet Protocol (IP) address, Internet and computer geolocation can be completed. The process of finding geolocation involves looking up an IP address on a WHOIS service and extracting the registrant's physical address. Information related to location like country of the device, region, city, postal/zip code of the device, latitude, longitude and timezone from where device was accessed are included about the IP address [6]. There can be need to identify where the web visitors are coming from to know where the potential customers are on the ecommerce website that will enable various enhancements on the website like to pre-populate country code on forms, display different language and reduce credit card fraud based on geographic location.

As per the proposed methodology, first the server log data is captured after which its essential attributes like ip address, time stamp & resource are extracted to get customer purchase related information. Also, the ip address that's extracted from server log is mapped to relevant geographic location by taking reference from sites like whatismyipaddress.com, thus extracting geolocation details. Then these details are fed to mining algorithm to find location based customer purchase pattern.

The remainder of the paper is structured as follows: Section 2 explains the motivation behind choosing this project, section 3 states background and literature review, section 4 explains the proposed approach and the last section discusses the results and observations.

2. MOTIVATION

There is overload of data that is created at a web server on each click of the client/visitor on the website. To track customer interest, their browsing information is stored on the web server log so that it can be made use of proficiently to make changes on the website design or to provide customer personalization, thus enhancing customer retention and increasing online business revenue. Furthermore, extracting the geographical location of the customer can suggest the purchase pattern pertaining to a particular geographical boundary, thus extending location based business scope.

The main motivation for this project is to improve business opportunities on internet using web usage mining by tracking customer behavior in order to provide web personalization. Web usage mining fundamentally has many advantages which makes this technology attractive as it has enabled personalized marketing for e-commerce, which eventually results in higher trade volumes. Thus companies can establish better customer relationship by understanding their needs better and reacting to them faster. They can increase productivity by target pricing on the basis of the profiles created. It is also possible to find the customer who might default to a competitor and to try to get interest of such customers back by providing them with promotional offers, thus reducing the risk of losing a customer or customers.

3. LITERATURE REVIEW

Data mining technology applied on World Wide Web is known as web mining. It is used to extract interesting and possibly useful patterns and unseen information from the web documents and through web activities. Web usage mining focuses on examining search logs or other activity logs to find interesting patterns of customers browsing behavior. Learning user profiles is one of the main applications of web usage mining [3].

Web data mining process is divided into the some stages as follows: source data collection, data preprocessing, data storage, pattern discovery and pattern analysis

1. Source data collection

Web server log files that are available on the Web server are the main source of data while performing mining data from web. Web log files have the history of the guests/customers browsing behavior. Web log files can be the server log, agent log and client log.

2. Data Preprocessing & Data Storage

Data from the server log have redundant & unwanted entries like error access records. To remove these from the web server log; its cleansing is performed thus generating meaningful data. Once processing is done, these data are stored in the database in order to get ready to be taken out and used. Web data mining uses transaction database more than relational database.

3. Pattern Discoveries

The next step of mining process is pattern discovery, which includes classification analysis, association rule discovery, sequential pattern discovery, clustering analysis and dependency modeling.

a) Classification analysis establishes categories of users by classifying data, as per the pre-defined categories on the basis of the user profile.

b) The association rule discovery is used for finding the relevant rules from the Web log database access information.

c) Sequential pattern discovery is to hit the model which has time series relations. On the websites' server logs, the user's sequence of behavior seems uneven due to odd visit of the user, thus it generates a discontinuous time series.

d) The clustering analysis classifies user or data items with similar type of personality and it collects these similar ones together. It can help with marketing decisions.
e) The dependency modeling is used to build up a model which can affirm the dependence between the different variables in Web field.
4. Pattern Analysis
A pattern set is generated by using an appropriate pattern discovery algorithm. Pattern analysis is primarily to select pattern we are concerned in from the pattern set. Its objective is to find out a worthy model, i.e., the rules and models we are interested in and then providing graphical user interface by using visualization techniques to users [3].
Business through e-commerce can be improved if location based data is available to the servers. Location based information further provides a unique opportunity to study customer activities through data analysis in a spatial-temporal-social perspective thus allowing a variety of location based services from marketing to disaster relief [7].

4. PROPOSED SYSTEM FLOW

The proposed system basically consists of 4 modules:

- Source data collection
- Data preprocessing & mapping location based data
- Pattern Discovery & analysis

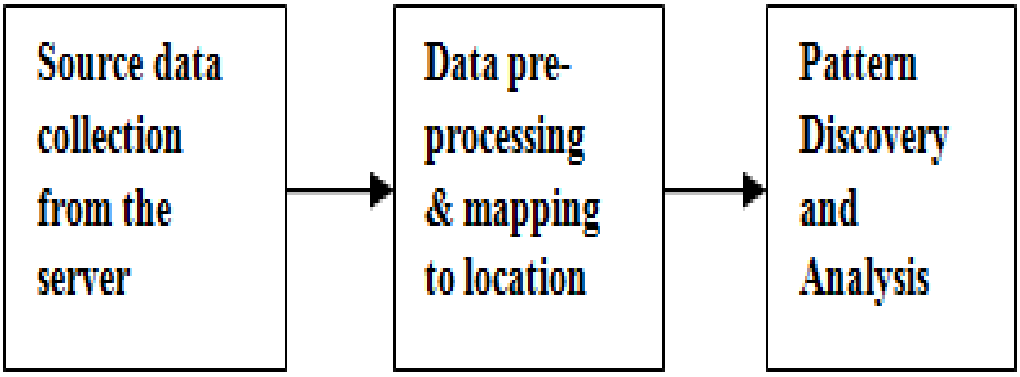

Figure 1: System Flow Diagram

Source data collection: As a first step towards implementation: We got transaction data from the website having vital information like ip address, time stamp & resource accessed. In mining of data from the web, server log files on the Web server are the main source of data. Web log files contain the history of the visitor's browsing behavior [8].
Data Preprocessing & mapping to location based information: The actual data collected have certain features such as redundancy, ambiguity and incompleteness, In order to mine knowledge more effectively, pre-processing the data collected is essential. Preprocessing can give accurate and concise data for data mining [3]. Also as per the proposed method mapping this cleaned data log to e-commerce related content is done by translating them to geolocation based information. Out of the many attributes pertaining to each entry of the server log, by considering only IP address, resource & timestamp customer behavior pattern can be mined. Going one step further, customers' location can be tracked by mapping IP address with IP look up support, in order to offer location based web personalization. This data will be stored in the database to get ready to be extracted.
Pattern Discovery & Analysis: Pattern discovery technique includes classification analysis, association rule discovery, sequential pattern discovery, clustering analysis and dependency modeling [9]. The association rule discovery is used for forming the relevant rules from the Web log and database access information. In e-commerce, association rules are mainly used for discovering which products customers usually buy as well when they buy something. Then companies can recommend customers some products that they may be interested in [10].

5. RESULT AND OBSERVATION

First step towards the proposed method is to get transaction data from the site having details like ip address, timestamp, resource.
Using an open source e-cart solution, an e-commerce shopping site has been developed & hosted to be able to perform transactions towards the realization of the proposed method - www.silspo.com/morrison

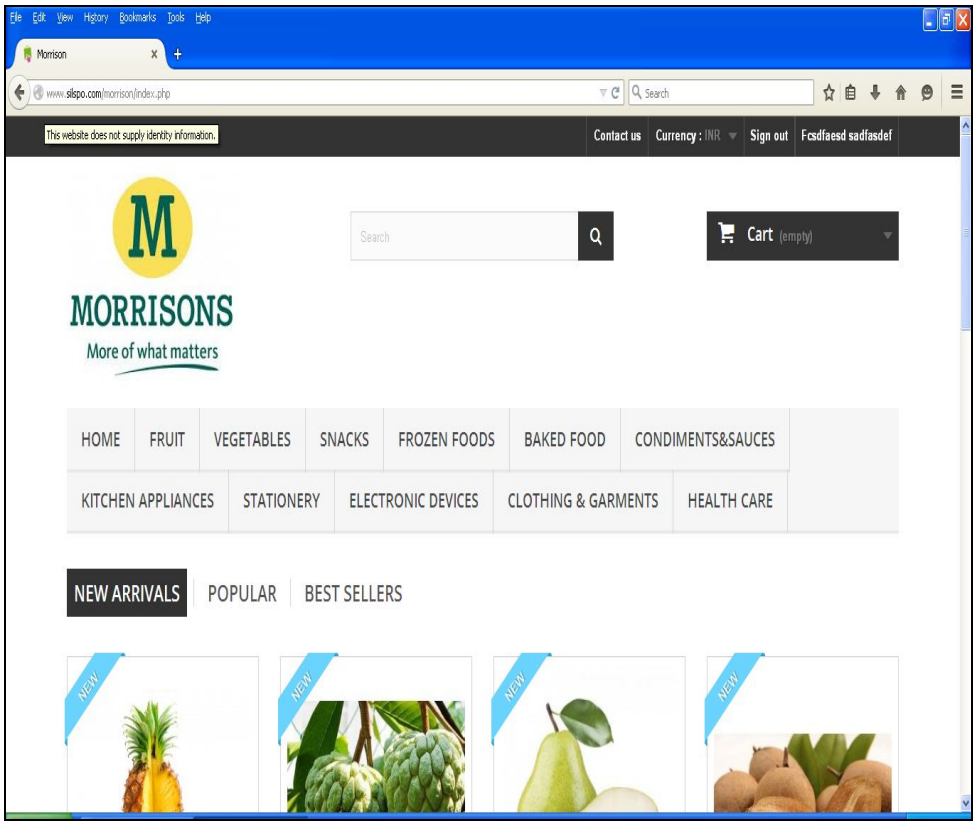

Figure 2: Screenshot of the site

Only the webmaster or administrator has access to the server side data like the web server log, server database, etc.

```
116.72.178.236 - - [08/Feb/2015:01:24:22 -0700] "POST /morrison/index.php?rand=1423383862551 HTTP/1.1" 200 138 "ht
103.230.222.134 - - [08/Feb/2015:01:24:53 -0700] "GET /morrison/index.php?id_category=22&controller=category HTTP/
103.230.222.134 - - [08/Feb/2015:01:24:54 -0700] "GET /morrison/themes/default-bootstrap/cache/v_2_9fe9bb65f1702ab
103.230.222.134 - - [08/Feb/2015:01:24:55 -0700] "GET /morrison/themes/default-bootstrap/cache/v_2_c7e513b8b25e953
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/4/7/47-small_default.jpg HTTP/1.1" 200 2856
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/4/5/45-small_default.jpg HTTP/1.1" 200 5071
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/5/5/55-small_default.jpg HTTP/1.1" 200 3838
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/en-default-home_default.jpg HTTP/1.1" 200 66
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/4/1/41-small_default.jpg HTTP/1.1" 200 2527
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/4/2/42-small_default.jpg HTTP/1.1" 200 5071
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/9/4/94-small_default.jpg HTTP/1.1" 200 2780
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/img/p/5/6/56-small_default.jpg HTTP/1.1" 200 3536
103.230.222.134 - - [08/Feb/2015:01:25:01 -0700] "GET /morrison/themes/default-bootstrap/img/pagination-li.gif HTT
```

Figure 3: Screenshot of the server log

After retrieving the web sever log, relevant information from this server log is transferred to the database. Also ip address is mapped to its equivalent location using appropriate ip lookup sites. The database table into which relevant information is stored is named as business relevant resource table. It has details like ipaddress, email id of customer, product name & category and also the location mapped from ip address.

id	ip_address	email	prod_name	prod_cat	order_date	location
126	150.107.183.74	divya.thakur@gmail.com	casual-shirts	clothing-garments	2015-09-04 19:50:33	Ulhasnagar
127	150.107.183.74	divya.thakur@gmail.com	capsicum	vegetables	2015-09-04 19:50:33	Ulhasnagar
128	150.107.183.74	divya.thakur@gmail.com	bourbon	snacks	2015-09-04 19:50:33	Ulhasnagar
129	117.218.24.91	prashanthegde010101@gm...	maskachaska	snacks	2015-09-05 19:40:50	Bangalore
130	117.218.24.91	prashanthegde010101@gm...	pickwick-wafers-s...	snacks	2015-09-05 19:40:50	Bangalore
131	117.218.24.91	prashanthegde010101@gm...	printer-paper	stationery	2015-09-05 19:40:50	Bangalore
132	59.94.36.222	nikhilbhosale1234@gmail.com	delhi-apple	fruit	2015-09-23 18:54:28	Mumbai
133	59.94.36.222	nikhilbhosale1234@gmail.com	black-grapes	fruit	2015-09-23 18:54:28	Mumbai
134	59.94.36.222	nikhilbhosale1234@gmail.com	mango	fruit	2015-09-23 18:54:28	Mumbai
135	59.94.36.222	nikhilbhosale1234@gmail.com	chikoo	fruit	2015-09-23 18:54:28	Mumbai

Figure 3: Screenshot of Business relevant resource table

In the next step, information from the database is redirected to a .arff file (attribute relation file format) that is later fed to mining algorithm.

```
@RELATION customer

@attribute prodname {shimla-apple,chicago-apple,
@attribute location {Bangalore,Dombivali,Mumbai,

@data
delhi-apple,Bangalore
sangli-watermelon,Bangalore
nutrichoice,Bangalore
cheese,Bangalore
veg-patties,Bangalore
burger-patties,Bangalore
cheese,Bangalore
micromax-4,Dombivali
long-beans,Dombivali
brinjal,Dombivali
tshirts,Dombivali
pants,Dombivali
formal-shirts,Dombivali
shirts,Dombivali
casual-shirts,Dombivali
bajaj-hand-blender,Dombivali
vicks,Dombivali
pain-relief-gel,Dombivali
ipad,Dombivali
```

Figure 3: Screenshot of .arff file

Finally the .arff file is passed to an appropriate mining algorithm –Apriori algorithm to find association rules between location and products.

6. CONCLUSION

The proposed system intends to provide geolocation based web mining by reflecting an association between products and location, thus suggesting which kinds of products are popular among customers of a particular location. Location based details are extracted from the IP address of the machine from where customer has accessed the site to perform the purchase of the product/products from the site.

The traditional methodology of performing web mining is implemented for web usage mining in order to extract customers' online behaviors and browsing patterns, but with an additional idea of getting customers' location to achieve location based web personalization further to improve ecommerce business trends and customer retention

```
Apriori
=======

Minimum support: 0.01 (1 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 103

Size of set of large itemsets L(2): 124

Best rules found:

 1. prodname=shimla-apple 11 ==> location=Ulhasnagar 11    conf:(1)
 2. prodname=washington-apple 3 ==> location=Ulhasnagar 3    conf:(1)
 3. prodname=lady-sfinger 2 ==> location=Mumbai 2    conf:(1)
 4. prodname=50-50 2 ==> location=Ulhasnagar 2    conf:(1)
 5. prodname=good-day-rich-butter 2 ==> location=Ulhasnagar 2    conf:(1)
 6. prodname=black-cumin 2 ==> location=Ulhasnagar 2    conf:(1)
 7. prodname=cinnamon 2 ==> location=Ulhasnagar 2    conf:(1)
 8. prodname=chicago-apple 1 ==> location=Ulhasnagar 1    conf:(1)
 9. prodname=fugi-apple 1 ==> location=Ulhasnagar 1    conf:(1)
10. prodname=mango 1 ==> location=Mumbai 1    conf:(1)
```

Figure 3: Screenshot of the output by apriori algorithm

REFERENCES

[1] http://en.wikipedia.org/wiki/Electronic_commerce (Last Visited 04/01/2015).
[2] http://en.wikipedia.org/wiki/Web_mining (Last Visited 24/08/2014)
[3] Mahendra Pratap Yadav, Mhd Feeroz, Vinod Kumar Yadav, “Mining the customer behavior using web usage mining In e-commerce”, IEEE-201S0, ICCCNT'12 26th_2Sdl July 2012, Coimbatore, India
[4] http://en.wikipedia.org/wiki/Server_log (Last Visited 09/01/2015)
[5] Web Usage Mining- Pattern discovery & its applications- Project research paper by Jinguang Liu & Roopa Datla
[6] http://en.wikipedia.org/wiki/Geolocation (Last Visited 10/01/2015)
[7] Huji Gao and Huan Liu, “Data Analysis on Location Based Social Networks”
[8] Yanduo Zhao, “The Review of Web Mining in E-commerce”, 2013 International Conference on Computational and Information Sciences.
[9] Sung-Shun Weng, Mei-Ju Liu, "Personalized product recommendation in e-commerce", IEEE International Conference on e-Technology, Service(EEE '04), 2004, pp.413-420, doi: 1O.1109IEEE.2004.1287340
[10] SUNEETHA, K. R. AND D. R. KRISHNAMOORTHI (2009). "IDENTIFYING USER BEHAVIOR BY ANALYZING WEB SERVER ACCESS LOG FILE." IJCSNS INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, VOL.9 NO.4, APRIL 2009