Demographic Analysis and Identification of E-Commerce Spending Tendencies

Naeem Th.Yousir
College of Information Engineering
Al-Nahrain University
Baghdad-Iraq
naeemms@yahoo.com

Abstract— E-commerce websites provide customers with a wide variety of navigational options and actions: users can freely move through different product categories, follow multiple navigational paths to visit a specific product or use different mechanisms to buy products, the user activities are stored in the web server log file. The Temporal Logic and model checking techniques are as a different to data mining techniques. These techniques have produced their applicability for open systems. The goal is to analyze the usage of e-commerce websites and to discover customers' complex behavioral patterns by means of checking temporal logic formulas describing such behaviors against the log model. At the beginning, web server logs are preprocessed to extract the detailed traces. When a user visits a product category or product sub category page, when a user adds a product to the wish list, when the search engine is used, etc. The business analyst can be used to the temporal logic patterns to formulate queries that could help users to discover and understand the way user use the website. Considering the website structure and contents as well as the different types of user's actions, these queries can check the existence of complex causality relationships between events contained in the user's sessions.

Index Terms— Weblog Preprocessing, Session identification, Demographic Analysis.

1 Introduction

The data mining technique is that this provides causal relations among events of a user trace, instead of providing with a global view of the whole session. Besides, it is the fact of avoiding the need of tagging the web pages. With respect to those approaches whose main objective is predicting the coming possible events, the approach allows having a global view of the sessions, making easier a global analysis of the user behavior, giving hints and facilitating the re-design of the website for a better adaptation to the user necessities [1]. An interesting feature of the approach followed that it properly fits the open nature of the use of e-commerce websites, where there are very few constraints for the users to navigate among site web pages. Another interesting feature of the followed mining approach is the fact of being able to analyze sequences of detailed events. The fact of considering the causal relations of events inside a user session, allowing to look for intra-session patterns can provide the analysts with a much more detailed perspective of a user behavior [2]. The tool considers an event as a complex entity, seen as the conjunction of a set of attributes. This allows not only having a detailed view of the user activities, but also a (hierarchical) view with multiple aspects (it is a matter of proposing different LTL formulas involving the desired attributes in which we are interested).

Demographics have some influence on if a person is online in the first place compared with the rest of the overall national population; they are, for example, more likely to be white and more highly educated. However, once people are online, whether they buy there and how much they spend has more to do with whether they like being online and whether the time they have for buying things elsewhere is limited [3]. However, even these general lifestyle characteristics explain only a small proportion of why people buy online and the amount of money they spend there. In the GVU data, demographics do show a slight influence: The higher a person's income, education, and age, the more likely that person will buy online, and the higher a person's income, the more online transactions that person is likely to make [4]. But this influence is barely significant; demographics alone predict 1.2% of decisions to buy or not buy and only 0.3% of the variance in the number of purchases made by online buyers. These results match findings from consumer behavior studies in other media in which demographics and lifestyle variables explain only a small percentage of people's choice behavior.

1.1 Main Architecture of the Project:

The User's information would separate multi-session patterns and correlate results with demographic information; while, online reviews would allow us to analyze customer's queries to recommend products. In demographic analyses, why men are so much more important for e-commerce retail, compared to in-store, and why they are more likely to buy on smartphones than women. Indexes online shopping spending by age group against the amount of time a given demographic spends online. Breaks down online spending habits of a teenager, including the brands and products they shop for. Examines the factors behind what drives online purchases among millennial.

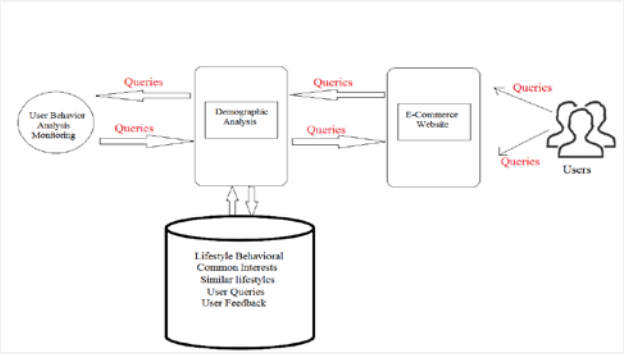


Figure 1: Main Architecture

2. Implementation Modules

2.1 Web Log Preprocessing

Web Log Preprocessing plays a major role in WUM. The raw log data doesn't suitable for Data Mining algorithms. Also, the size of the data is huge and demands memory and processor. The Web Log Preprocessing should produce reliable and good quality data so that the Data Mining algorithms can produce useful patterns. The Web Log Preprocessing algorithms should result in meaningful user sessions which are then analyzed by the Data Mining Algorithms. According to, weblogs can predict user's next request without disturbing them. However, the not all details/files available in web logs are appropriate for mining navigation patterns [5] [6]. So, the information from web logs needs cleaning before it can be used for prediction. Therefore, unnecessary product is deleted using cleaning algorithm. In several data preparation techniques used to improve the performance of the data preprocessing to identify the unique sessions and unique users is presented [7]. A Field extraction algorithm to separate the fields from the single line of the log file and a Data cleaning algorithm to eliminate inconsistent or unnecessary items in the analyzed data.

2.2 User session identification

The sequence of activities performed by a user from the moment he enters the website to the moment we leave the website is referred to as a session. In User Session identification, the web access log of every user is split into sessions and these sessions are further analyzed. Two methods based on time and navigation is usually considered for session identification. The time-oriented rely on session duration or page stay time. The navigation-oriented is based on user navigation pattern. Each method scans the user activity logs and divides the user activity into sessions. Session-duration based method: The total duration of a session cannot exceed a threshold Θ. A new request is considered as a new session if this request was given after the above threshold. Page-stay-time based method: Here, a threshold is employed as the maximum page stay time. If the current request was done within the above threshold then this request is added to constructed session.

Figure 2: session identification

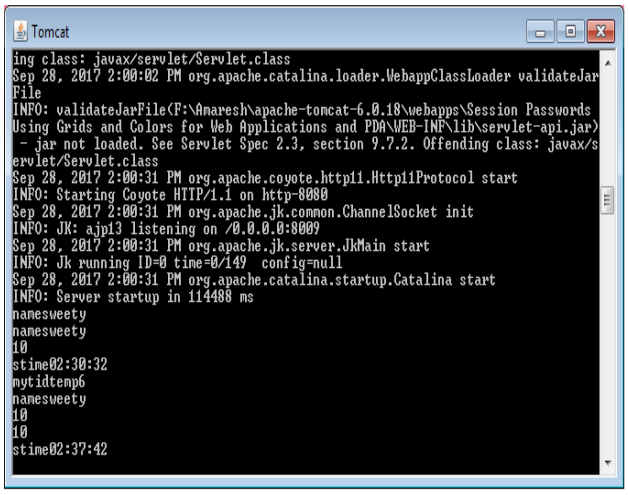
URL Name	Start Time	End Time	Session(min)
www.sony.com	02:43:04	03:29:49	46
www.sony.com	02:37:42	02:37:55	0
www.sony.com	02:30:32	02:31:02	1
www.sony.com	02:42:35	02:43:03	1
www.samsung.com	02:38:16	02:39:19	1
www.sony.com	02:34:35	02:35:42	1
www.samsung.com	02:35:43	02:36:33	1
www.samsung.com	01:22:27	01:23:39	1
www.sony.com	01:23:39	01:23:50	0
www.lenovo.com	01:10:57	01:14:07	4
www.micromax.com	01:14:07	01:16:00	2
www.lenovo.com	01:16:01	01:18:11	2
www.sony.com	01:18:11	01:20:06	2

Figure 3: Spending Tendency

2.3 Website Analysis

Users of any online product sales site navigate through the different web pages executing two types of interactions: either a GET operation means to retrieve some information or a POST operation, usually requesting the website to execute some action, such as adding some product to the cart, buying some product, logging in, etc. The website log records such actions together with some associated information, such as the IP the user is connected from or the time at which the interaction occurs, for instance. Some of these actions correspond to events that are common to any e-commerce website such as the ones related to visiting the sections containing products. Therefore, a general way of classifying the events in the weblogs according to the product categorization can be proposed. From now on, we are going to describe the proposed approach to relate the website structure and the events in the log, to identify a meaningful set of events, and to ask for behavioral usage patterns using model checking based on the previous classification.

Figure 4: Website Analysis



3 ALGORITHM IMPLEMENTATIONS

Association Rule:

In the conditions of web usage mining, once sessions have been identified in the association rules can be used to relate pages that are most often referenced together in a single server session. Such conditions mention the possible relationship between pages that are often looking at together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Since usually, such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to support for items under consideration.

C->D [Support, Confidence]

This means the presence of item C leads to the presence of item D, with [Support]% occurrence of [C, D] in the whole database, and [Confidence]% occurrence of [D] in a set of records where [C] occurred.

$$\text{Support}=P(A\cap B) = \frac{\textit{number of sessions that contain A and B}}{\textit{total number of sessions}} \quad (1)$$

$$\text{Confidence}(C\text{->}D) = \frac{support(C\cap D)}{support(C)}$$

Collaborative Filtering:

Collaborative filtering method based on the standard cosine similarity or Pearson correlation to compute the similarity between two users [8][9]. For arbitrary users and, the number of common product shared by them can be defined as

$$C_{ij}=\sum_{l=1}^{m} ali\ alj \quad (2)$$

Generally, for standard cosine similarity computation, let **Sij** denote the similarity between **ui** and **uj** and let **k(ui)/k(uj)**denote the degree of the user **ui/uj**; namely, how many objects are collected by this user? So, we can formulate the expression as

$$S_{ij}=\frac{cij}{\sqrt{k(ui)k(uj)}}=\frac{\sum_{l=1}^{n} ali\ alj}{\sqrt{k(ui)k(uj)}} \quad (3)$$

Algorithmic Frame:

```
Algorithm CF-M: Calculating the similarity between users
Begin
Get A
n=size(A,2),m=size(A,1);
parameter ar;
preference degree v();
range of the rating score M;
S=Zeros(m,m),o=sum(A),u=sum(A');
  For i=1: m
    For j=1 : m
      X=(u(i)*(u(j)))^(-0.5);
          For z=1 : n
              Y=y+a(i,z)*a(j,z)*((1-abs(v(i,z)-
v(j,z))/M)/o(z))^ar
          End
          S(i,j)=x*y;
          X,y=0;
    End
  End
End
```

4 CODE IMPLEMENTATION

4.1 Session identification

```
//date
        java.util.Date now = new java.util.Date();
        String date=now.toString();
         String DATE_FORMAT = "dd-MM-yyyy";
         SimpleDateFormat sdf = new SimpleDateFor-
mat(DATE_FORMAT);
        String strDateNew = sdf.format(now) ;
//time
        String DATE_FORMAT1 = "hh:mm:ss";
        SimpleDateFormat sdf1 = new SimpleDateFor-
mat(DATE_FORMAT1);
        String strDateNew1 = sdf1.format(now) ;
        String etime="00:00:00";
        int time=0;
```

4.2 Website Analysis

```
<%
                    Connection con4=null;
                    Statement st4 = null;
                    ResultSet rs4 = null;

                    try{
                    con4=databasecon.getconnection();
                    st4 = con4.createStatement();
                    String qry4 ="select * from website
where urlname!='"+sname+"' and category='"+mycategory+"'
order by count DESC";
                    rs4 = st4.executeQuery(qry4);
                    while(rs4.next()){%>
                    <li id="category-active"><a
href="up5.jsp?search=<%=rs4.getString("urlname")%>"><%=r
s4.getString("urlname")%></a></li>
      <%}
                    }

                catch(Exception ex4){
                 out.println(ex4);

              }
```

5 Result

The customer is the ultimate critic who decides the fate of the market players, and there are several factors that influence the activities of the customer in the market. The customer is influenced by several factors like demographic, social, economic, cultural, political and technological factors; demographic factors impact customer's lifestyle and play a major role in determining purchase decisions like age, gender, education, occupation, income level.

Table 1: Demographic Analysis based on Age

Demographic Factors	Spending tendency(min)
15-20	30
20-30	35
30-40	25
40-50	15
>50	10

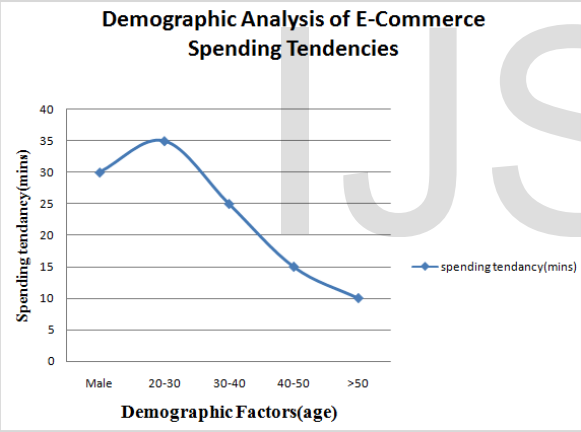


Figure 5: Demographic Analysis based on Age

Table 2: Demographic Analysis based on Gender

Demographic Factors	Spending tendency(min)
Male	100
Female	70

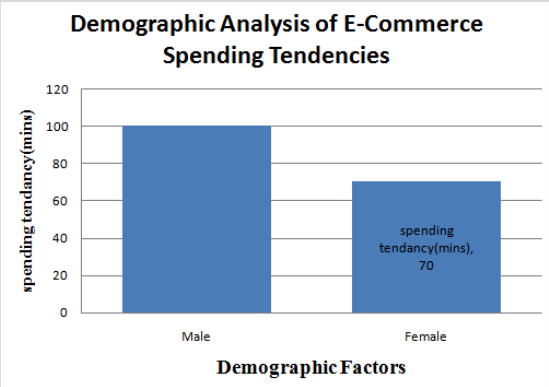


Figure 6: Demographic Analysis based on Gender

6 CONCLUSIONS

The market is a huge area where every seller can find a place if he provides value and be trustworthy to the customer. Convenience in navigating e-tailor's website, quick loading and an accurate product/service delivery system is considered essential for customer's acceptance of online retailer. Further studies can be conducted to know the impact of other socio-psychological factors on online buying behavior intention and acceptance levels by customers. Also, the Preprocessing of data is an essential activity which will help to improve the quality of the data and successively enhance the quality of mining results, as a result it enhances the performance of the system. Web Log Preprocessing is one of the important steps in Web Usage Mining. Data Cleaning step revealed that a major portion of Web Log usually consists of irrelevant and redundant data which must be eliminated to speed up the upcoming mining process.

7 Future Scope

Currently, only a simple neural network architecture has been employed for user and product embeddings learning. In the future, more advanced deep learning models such as Convolution Neural Networks can be explored for feature learning. We will also consider improving the current feature mapping method through ideas in transferring learning. We also plan to extend the set of studied patterns to analyze more behavioral patterns and to facilitate their automatic discovery. For that, a side-by-side work with specialists of the problem domain is required to define a set of interesting queries as wide as possible. Additionally, extending the web server logs with information about users or online customer reviews is going to be studied.

References

[1] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications." Hingham, MA, USA: Kluwer Academic Publishers, Jan. 2001, vol. 5, no. 1-2, pp. 115–153.

[2] W. W. Moe and P. S. Fader, "Dynamic conversion behavior at ecommerce sites," Management Science, vol. 50, no. 3, pp. 326–335, 2004.

[3] Y. S. Kim and B.-J. Yum, "Recommender system based on click stream data using association rule mining," Expert Systems with Applications, vol. 38, no. 10, pp. 13 320–13 327, 2011.

[4] F. M. Facca and P. L. Lanzi, "Mining interesting knowledge from weblogs: a survey," Data & Knowledge Engineering, vol. 53, no. 3, pp. 225–241, 2005.

[5] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. M. Prez, and I. Perona, "Web usage and content mining to extract knowledge for modelling the users of the bidasoa turismo website and to adapt it." Expert Syst. Appl., vol. 40, no. 18, pp. 7478–7491, 2013.

[6] Y. H. Cho and J. K. Kim, "Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce," Expert Systems with Applications, vol. 26, no. 2, pp. 233 – 246, 2004.

[7] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining:

Discovery and applications of usage patterns from web data," SIGKDD Explor. Newsl., vol. 1, no. 2, pp. 12–23, Jan. 2000.

[8] B. Singh and H. K. Singh, "Web data mining research: a survey," in Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on. IEEE, 2010, pp. 1–10.

[9] F. M. Maggi, R. P. J. C. Bose, and W. M. P. van der Aalst, Efficient Discovery of Understandable Declarative Process Models from Event Logs. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 270– 285.