A Web Usage Mining for Modeling Buying Behavior at a Web Store using Network Analysis

Miseon Lim¹, Hyunsoo Byun² and Jinhwa Kim¹*

¹School of Business, Sogang University, South Korea; jinhwakim@sogang.ac.kr
²Department of Public Management Information Systems, Korea National University of Transportation, South Korea

Abstract

Understanding visitors’ invisible behaviors and responding with appropriate answers are important issues in continually increasing online market. To promote online transactions, customers’ behavior should be predicted correctly to keep low purchase conversion rate. In this study, we suggest an approach based on the idea that customers’ sessions in a web store can be transformed into the structure of a graph, which are represented as density of a session based on a graph theory. Online users visit lots of sites and their activities include information acquisition and browsing. The history of these activities can be used to construct a relationship network among web sites. This study analyzes this visit history made by website visitors with graph theory. The density of a network refers to the differentiated degrees of relationship among objects. In this study, we dichotomize into “purchase” and “no purchase group” since predicting whether a customer will buy or not buy our products is an important issue in web stores. We collect data on sessions which are a sequence of page views or a period of sustained web browsing. We model the sessions on the basis of density of a graph, which resulted in DOS (Density of a Session). The performance of other predictors including DOS is compared to that of suggested method in this study. Predictors are TVT (Total Visit Time during a period of a visit), AVT (The Average Time per Page Viewed), TNC (Total Number of Clicks), TPP (Total Number of Product-Related Pages Viewed), and DOS (Density of a Session Based on Graph Analysis). The study found that all predictors except total visit time are useful to differentiate between “purchase” and “no purchase” group. And we conducted Logit Analysis to examine the performance of each purchase prediction method. The results from Logit Analysis show that DOS predicts purchase behavior better in comparison with other predictors. It means understanding customers’ sessions with respect to a graph structure is useful to predict whether a customer will buy or not buy products in a web store.

Keywords: Click Stream Data, Predicting Customer Behavior, Web Usage Mining

1. Introduction

Many companies, regardless of their size and types, manage on-line channels alone with their traditional outlets. Web stores, also known as on-line virtual stores or on-line shopping malls, are expected to be new distributors in the market, especially for B2C retailers. But conversion rates, which mean how many visitors actually end up browsing a web store by purchasing a product, is very low. Conversion rates of web stores are mostly under one-digit percent¹⁰⁻¹⁹,2₄,2₈⁻₃⁰. Exceptionally, Bucklin et al.* showed 22.9% because their on-line store is specialized in selling wine, and only log-in users can enter the web site. Under these circumstances, predicting and understanding customer behavior in online is useful for the web stores.

Consumer behaviors in the Internet are dynamic and have distinct features. The behaviors of visitors in internet stores can be categorized into several types; buying, searching, browsing, and learning⁵,⁶. Though Internet is a virtual space, there are efforts in capturing unseen behavior of internet users in many ways. Lots of researches have paid attention to clickstream data. Clickstream data is also called web log data, which is the history of web users’ behavior in web sites. Clickstream data provides not only the information with regard to
the purchase behavior, but it also gives the information
concerning the trajectory at web stores\textsuperscript{14}.

Customers in web traverse sites on many products and
their travel history can be represented with relationships
among products. The graph theory is useful to analyze
these constructed relationships. The graph theory is
applied to represent these relationships and it is used
to solve combinatorial problems. It is also used to solve
problems such as managing large-sized data warehouses
and clustering categorical data\textsuperscript{2}. Based on graph theory,
objects have relationships in a decision space. The
relationship among these objects can be applied at a
recommending system or market basket analysis\textsuperscript{4}.

2. Literature Review

Shopping behaviors on on-line can be divided into
exploratory searches and goal-directed searches\textsuperscript{13}.
Moe\textsuperscript{17} extends this typology to four shopping behaviors
(directed buying, hedonic browsing, search/deliberation,
knowledge building). The amount of time they spend
looking for product information and the amount of
time they spend choosing from a list of products are
major factors influencing online buying behavior\textsuperscript{3}.
Many people use the Internet for checking e-mails and do their
business at home or workplace. They access product and
service information directly or indirectly. Therefore it is
an important problem that estimating how much time
they spend in searching for product information and two
much time they spend in buying products actually. As it
becomes faster and more convenient to find information
and buy products in on-line store, the amount of
time they spend is a very important factor influencing users’
satisfaction on on-line services. Bellman et al.\textsuperscript{3}
shows that
the wired lifestyle and time starvation influence online
purchases mostly. And they show that demographic
information such as ages and genders shows a little
influence. Emmanoulides and Hammond\textsuperscript{11} also show that
demographic variables such as ages, genders, and income,
which are expected to determine Internet usage, do not
have much influence on Internet usage. So Bellman et al.\textsuperscript{3}
claimed that convenience, customization, and ordering
processes should be considered to be major factors in
designing web services. The similar results are shown in
the panel data analysis by Lohse et al.\textsuperscript{15} that the longer
the amount of time spent in on-line store, the greater the
probability of purchasing.

Moe and Fader\textsuperscript{18} focus the relationship between
visiting and purchasing behaviors. They developed a
model to understand purchase conversion of customers
and visit patterns. Sismeiro and Bucklin\textsuperscript{25} decompose
online ordering of the website into sequential Nominal
User Tasks (NUT) and examine purchase behaviors.
Montgomery et al.\textsuperscript{19} show that users’ initial travel paths in
web reflect visitors’ goal. They apply this model to predict
purchase conversion and suggest a method to predict
purchase decision by users. Others classify predictors
into four types. These four types are general clickstream,
detailed clickstream, customer demographics, and
historical purchase behavior. These variables help predict
purchase decision.

3. Network Analysis at Web
Stores

The decision space in this study is an aggregate of
networks linked together, and many objects in a web
site are connected each other. There are associations and
relationships in a decision space, which can be presented
in the form of adjacent data. If item \(a\) and item \(b\) are
purchased at the same time, then it can be considered
to be adjacent between them. This adjacency concept
can be applied to products recommendation or objects
visualization in web sites\textsuperscript{2,9}. Following graph theory, graph
and its matrix represent relationships among objects.

\[
\begin{array}{cccccc}
C_1 & P_1 & P_2 & C_2 & P_3 & P_4 \\
C_1 & 0 & 1 & 0 & 1 & 0 & 0 \\
P_1 & 0 & 0 & 1 & 0 & 0 & 0 \\
P_2 & 0 & 1 & 0 & 0 & 0 & 0 \\
C_2 & 1 & 0 & 0 & 1 & 0 & 0 \\
P_3 & 0 & 0 & 0 & 0 & 1 & 0 \\
P_4 & 0 & 0 & 0 & 1 & 0 & 1 \\
\end{array}
\]

Figure 1. An example of an adjacency matrix and a
directed graph of a session in a web store\((C_n: \text{Category } n, P_m: \text{Product } m)\)

A graph consists of two components, vertex and arc.
The graph is a set of adjacent data and can be transformed
into an adjacency matrix. An adjacency matrix is an \(n \times n\)
square matrix, \(A = (a_{ij})\), where \(a_{ij} = 1\) if item \(i\) is adjacent
to item \(j\), and \(a_{ij} = 0\) otherwise. The visually represented
graph and the adjacency matrix of it can be applied
usefully for network analysis. A graph consists of points
(vertex) and lines (arc). The point means an object and the
line means a relation. The types of graph are divided into undirected and directed graphs according to directness. In this study, we selected directed graphs because of orders and sequences in a website. Also, it is needed to consider the value of relations. The value means the intensity of relations. For example, page views, frequency, and duration et al.

On-line customers move sequentially from item to item in web stores. Also most web stores arrange categories and products. A visitor in the web store moves among categories or products and this eventually constructs a session. For example, this visitor’s movement can be represented as in Figure 1.

The structure of network in a session is needed to analyze quantitatively. We use density of a graph for modeling. The density of a network means the degree linked among objects and is determined by how many the objects connected and the degree to which one object connected to the others. The density represents the degree connected actually by maximum linkage, which can be connected among objects\(s^{2,22,26}\). The density of a network can be represented as in Equation 1.

\[
D = \frac{l}{p(p-1)/2}
\]

\(D\): density  
\(l\): the number of lines in a network  
\(p\): the number of points in a network

The denominator \(p(p-1)/2\) is the most number of relation which can exist in this network, and the range of density \(D\) is between 0 ~ 1. But Equation 1 is the case of undirected graph. At the case of Directed graph, the relation \(A \rightarrow B\) and \(B \rightarrow A\) are different, so the maximum number of relation is \(p(p-1)\). The network density of directed graph can be represented like Equation 2.

\[
D = \frac{l}{p(p-1)/2}
\]

\(D\): density  
\(l\): the number of lines in a network  
\(p\): the number of points in a network

Except directed or undirected graph, there are valued graphs to which relation value added. We focused on visiting time among properties of categories or products. Time is a major factor influencing online user behavior\(^3,15\). There are several types of time in detail. For example, the average time of visit, total visit time, duration time. Time is thought to predict online buying in many ways.

In this study, we propose the value of a relation using duration time, the total time spent viewing a product or a category. Two points \(a\) and \(b\), which have a relationship in a graph indicate duration time \(t_a, t_b\), respectively. A pair of \((a, b)\) represents duration time \(t_a + t_b\). Thus the value of a pair of two points means sharing time put together each duration time. Given a session (a visitor’s trajectory path), we assume the density of a valued graph based on it as DOS (Density of a Session).

High density in a network means the degree of linkage among objects is concrete, which means delivery or proliferation of information of an object is dynamic and fast. Therefore, involvement or interest of products is high and purchase possibility increases.

\[
DOS_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (T_{ik} + T_{jk})}{N(N-1)}
\]

\(DOS_k\): density of a session \(k\)  
\(N\): the number of nodes in a session \(k\)  
\(T_{ik}\): duration time point \(i\) in a session \(k\)

We selected a certain category and products in clickstream data. We constructed adjacency matrix for each purchase (not purchase) group according to whether they buy products or not (Figure 2). In the representation of relationships among items and web pages, blank means no relationship and 1 means relationship among them. There are more movements or links in “not purchase” group than those in "purchase" group. Thus, customers in “no purchase” group generally have more search and information acquisition time.

The number of movements, represented in dot, in “purchase” group are smaller compared to “not purchase” group, which means customers in “purchase” group have high intention to buy as they focus a certain product to buy. In this way, understanding different networks quantitatively, according to whether buy or not, can be used predicting buying behavior of customers in web stores.
A Web Usage Mining for Modeling Buying Behavior at a Web Store using Network Analysis

4. Results

4.1 Data Collection and Processing

Clickstream data is a type of web log data and encapsulates visitors' behavioral history. Clickstream dataset is shown in the form of URL viewed in Location Bar of web browser (ex. Internet Explorer). This dataset is very large and redundant in many cases. In general, it has been focused on “purchase” data set traditionally. Clickstream data, however, holds much of “not purchase” data. It is also difficult to manipulate the datasets because the data is very large and complicated with many components.

There are four methods to collect clickstream data. Firstly, retailer data is from log files in web servers of online vendors. Secondly, shop bot data is user's trajectory data, intermediating consumer's movement. Thirdly, experimental data is to purpose to analyze the data through factitious experimental design. Finally, panel data is collected by Internet survey from companies using software embedded in their PC.

We acquired and analyzed clickstream data of a web store to apply our model to the behavior of visitors. The company is a small sized web store, which sells desktops, laptops, and related products. The web store is one of the typical online companies. We collected the dataset of visitors' behavior from September 1, 2005 to January 31, 2006 from the web store. Data preprocessing was conducted to construct more significant dataset. Clickstream dataset includes various objects such as hyperlink, images, movies and so on. Only URLs are related to categories and products. A page view or page visit refers to a user's request with mouse clicking or actual input of URL in the user's web browser. In many cases users may click back button of browser to return to the former screen. This action is excluded as a page view or page visit. A session means a sequence of page views or a period of sustained web browsing. A session includes a visitor's whole objects viewed entering a web store. Because a session becomes a basis for graph analysis, our modeling is focused on density of a session. Duration time (or visit time) is defined as the time from a page viewing to next page.

During the monitoring period, the number of visitors is 598 and they had 5,967 sessions. Total number of purchase sessions is 247 sessions, which produces a conversion rate 4.1%. After refining the data set, 1,118 sessions were selected out of 5,967 sessions, where purchase sessions were 190. The final dataset includes 1,118 sessions, which include purchase sessions of 17%.

4.2 Comparison between Purchase and No Purchase Group

Online user groups are divided into some types according to viewpoints or issues. In this study, we dichotomize into two groups; “purchase” group and “not purchase” group. Buying or not is an important issue because of very low purchase conversion rate in online transactions. In fact, a small part of customers incurs a large part of total profits, which means that most of the customers are unprofitable. We classify the user groups according to whether buy or not. We propose a model to predict purchase decision using density of a session from each group.

Padmanabhan et al. shows that visit time spent in a certain session in a web site is a significant factor to predict possibility of a purchase. Bucklin and Sismeiro explain that total visit time spent to view a specific web page determined consistent web browsing. Emmanouilides and Hammond suggest a model to predict visitors' frequency of use and active or lapsed degree of users. Li et al. develop a model to predict the number of web pages in a particular category viewed by a user in a single session of a website. Based on prior works, we assorted predictors including DOS (Density of a Session) and compare performance of them in Table 1.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVT</td>
<td>Total Visit Time during a period of a visit</td>
</tr>
<tr>
<td>AVT</td>
<td>The Average Time per page viewed</td>
</tr>
<tr>
<td>TNC</td>
<td>Total Number of Clicks</td>
</tr>
<tr>
<td>TPP</td>
<td>Total number of Product-related Pages viewed</td>
</tr>
<tr>
<td>DOS</td>
<td>Density of a Session based on graph analysis</td>
</tr>
</tbody>
</table>
Firstly, we conducted t-test to examine if there is any significant difference between “purchase” and “not purchase” group. The result is shown in Table 2. Predictors except TVT are significant between two groups. Difference in means of total visit time between two groups is less significant compared to that of other predictors. The p-value of TVT is 0.140, which means it is difficult to differentiate two groups with total visit time during a period of a visit in our study.

We divide 1,118 sessions randomly into training dataset and test dataset. The number of each dataset is the same (training set: 559, test set: 559). The ratio of two groups in each dataset is settled in accordance of the conversion rate of final dataset. The rate of purchase sessions of each dataset is 17%. To get reliable performances, we tried ten cross-validations. We conducted Logit Analysis to compare predictors to predict purchases. In Logit Analysis, the targeted variable is binary (buy or not).

\[
P_i = \frac{1}{1 + e^{-b_j X_{ij}}} \tag{4}
\]

The final output from a customer’s visit is shown as binary, whether to buy or not. A probability \(P_i\) means the \textit{a posteriori} probability of purchase by a session \(i\), of which range is from 0 to 1. \(b_j\) is referred to a estimated parameter, \(X_{ij}\) is referred to independent variable \(j\) for a session \(i\). Thus, \(P_i\) is defined as the “purchase” probability according to a certain variable of each type of predictors.

4.3 Results

Logit model is used to examine the significance of each predictor. Since all predictors are used to predict purchase respectively, they are not entered at a time but entered individually as in Table 4. If the parameter estimates are positive, the higher the predictor, the higher the probability of purchase. On the contrary, if the parameter estimates are negative, the higher the predictor, the lower the probability of purchase. The concept behind TVT and AVT is that the longer the time, the higher the probability of purchase. The other way, for TNC and TPP, if the number of frequency of clicks or pages viewed is increasing, the probability of purchase is decreasing. Only TVT does not differentiate between “purchase” and “not purchase” group. It is difficult to predict purchase with the length time spent in a website. When observing the result of Logit Analysis, the odds ratio is useful to analyze. If the odds ratio is higher than 1, the probability of the

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Group type</th>
<th>Number of sessions</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Levene's Test for Equality of Variances</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td>Sig.</td>
<td>t</td>
</tr>
<tr>
<td>TVT</td>
<td>no purchase</td>
<td>928</td>
<td>52.63</td>
<td>25.12</td>
<td>.37</td>
<td>.545</td>
<td>-1.08</td>
</tr>
<tr>
<td></td>
<td>Purchase</td>
<td>190</td>
<td>54.80</td>
<td>24.11</td>
<td>.545</td>
<td>1.08</td>
<td>-.08</td>
</tr>
<tr>
<td>AVT</td>
<td>no purchase</td>
<td>928</td>
<td>8.06</td>
<td>3.71</td>
<td>14.67</td>
<td>.000</td>
<td>-6.72</td>
</tr>
<tr>
<td></td>
<td>Purchase</td>
<td>190</td>
<td>10.39</td>
<td>4.74</td>
<td>.000</td>
<td>6.08</td>
<td>.000</td>
</tr>
<tr>
<td>TNC</td>
<td>no purchase</td>
<td>928</td>
<td>6.95</td>
<td>2.77</td>
<td>11.52</td>
<td>.001</td>
<td>6.09</td>
</tr>
<tr>
<td></td>
<td>Purchase</td>
<td>190</td>
<td>5.71</td>
<td>2.22</td>
<td>.001</td>
<td>3.43</td>
<td>.000</td>
</tr>
<tr>
<td>TPP</td>
<td>no purchase</td>
<td>928</td>
<td>3.65</td>
<td>1.70</td>
<td>21.44</td>
<td>.000</td>
<td>-6.71</td>
</tr>
<tr>
<td></td>
<td>Purchase</td>
<td>190</td>
<td>3.12</td>
<td>1.26</td>
<td>.000</td>
<td>4.35</td>
<td>.000</td>
</tr>
<tr>
<td>DOS</td>
<td>no purchase</td>
<td>928</td>
<td>4.23</td>
<td>3.51</td>
<td>32.16</td>
<td>.000</td>
<td>-10.53</td>
</tr>
<tr>
<td></td>
<td>Purchase</td>
<td>190</td>
<td>8.33</td>
<td>5.76</td>
<td>.000</td>
<td>-10.51</td>
<td>.000</td>
</tr>
</tbody>
</table>

Table 3. Results from Logit Analysis according to each predictor

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Wald (X^2)</th>
<th>DF</th>
<th>P</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>TVT</td>
<td>.004</td>
<td>.003</td>
<td>1.169</td>
<td>1</td>
<td>.280</td>
<td>1.004</td>
</tr>
<tr>
<td>AVT</td>
<td>.135</td>
<td>.022</td>
<td>38.418</td>
<td>1</td>
<td>.000</td>
<td>1.145</td>
</tr>
<tr>
<td>TNC</td>
<td>-.200</td>
<td>.035</td>
<td>32.730</td>
<td>1</td>
<td>.000</td>
<td>.819</td>
</tr>
<tr>
<td>TPP</td>
<td>-.242</td>
<td>.058</td>
<td>17.501</td>
<td>1</td>
<td>.000</td>
<td>.785</td>
</tr>
<tr>
<td>DOS</td>
<td>.272</td>
<td>.030</td>
<td>79.735</td>
<td>1</td>
<td>.000</td>
<td>1.312</td>
</tr>
</tbody>
</table>
event is high, which means the probability of a purchase is high. On the other hand, if the odds ratio is more lower than 1, the probability of the event is low, which means the probability of a no purchase is high. When odds ratio of AVT is 1.145 and if average time per page viewed increases by one second, the probability of “purchase” is 14.5% higher.

The result from purchase prediction is shown in Table 4, next page. Total sessions are randomly divided into training and test dataset at the portion of 50:50. The proportion of purchase group in each dataset is 17% alike.

As in Table 4, the performance of purchase prediction of DOS is consistently better than that of other predictors. To show more reliable performance comparison, we conduct 10 cross-validations. As the results of average performance of predictors from ten cross validations show that purchase prediction of DOS is the best, it can be used to predict purchase behavior of customers in a web store as a significant predictor.

<table>
<thead>
<tr>
<th>Cross Validation</th>
<th>Sample</th>
<th>TVT</th>
<th>AVT</th>
<th>TNC</th>
<th>TPP</th>
<th>DOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>train</td>
<td>49.8</td>
<td>58.2</td>
<td>60.2</td>
<td>52.2</td>
<td>64.5</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>52.5</td>
<td>65.0</td>
<td>61.1</td>
<td>60.4</td>
<td>76.5</td>
</tr>
<tr>
<td>2</td>
<td>train</td>
<td>48.8</td>
<td>60.9</td>
<td>58.2</td>
<td>54.8</td>
<td>68.2</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>53.1</td>
<td>61.7</td>
<td>63.0</td>
<td>57.8</td>
<td>72.8</td>
</tr>
<tr>
<td>3</td>
<td>train</td>
<td>57.5</td>
<td>65.6</td>
<td>58.2</td>
<td>56.9</td>
<td>72.1</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>54.4</td>
<td>59.9</td>
<td>62.9</td>
<td>56.0</td>
<td>68.4</td>
</tr>
<tr>
<td>4</td>
<td>train</td>
<td>48.5</td>
<td>62.2</td>
<td>58.3</td>
<td>54.7</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>51.5</td>
<td>62.5</td>
<td>63.2</td>
<td>57.9</td>
<td>73.8</td>
</tr>
<tr>
<td>5</td>
<td>train</td>
<td>51.8</td>
<td>61.2</td>
<td>56.9</td>
<td>52.8</td>
<td>69.9</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>48.8</td>
<td>62.0</td>
<td>64.4</td>
<td>59.7</td>
<td>71.2</td>
</tr>
<tr>
<td>6</td>
<td>train</td>
<td>50.2</td>
<td>60.5</td>
<td>65.2</td>
<td>59.5</td>
<td>72.5</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>52.1</td>
<td>62.0</td>
<td>56.1</td>
<td>53.1</td>
<td>68.3</td>
</tr>
<tr>
<td>7</td>
<td>train</td>
<td>58.4</td>
<td>56.4</td>
<td>65.0</td>
<td>57.4</td>
<td>67.3</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>60.5</td>
<td>66.9</td>
<td>56.2</td>
<td>55.2</td>
<td>73.8</td>
</tr>
<tr>
<td>8</td>
<td>train</td>
<td>63.2</td>
<td>67.9</td>
<td>54.8</td>
<td>53.2</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>61.1</td>
<td>56.4</td>
<td>66.3</td>
<td>59.4</td>
<td>65.7</td>
</tr>
<tr>
<td>9</td>
<td>train</td>
<td>55.1</td>
<td>63.7</td>
<td>55.1</td>
<td>51.2</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>51.5</td>
<td>60.5</td>
<td>66.2</td>
<td>61.5</td>
<td>70.8</td>
</tr>
<tr>
<td>10</td>
<td>train</td>
<td>49.2</td>
<td>63.5</td>
<td>63.2</td>
<td>61.9</td>
<td>75.8</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>52.2</td>
<td>59.5</td>
<td>58.2</td>
<td>50.5</td>
<td>65.6</td>
</tr>
<tr>
<td>Average of 10 trials</td>
<td>train</td>
<td>53.3</td>
<td>62.0</td>
<td>59.5</td>
<td>55.5</td>
<td>70.2</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>53.8</td>
<td>61.6</td>
<td>61.8</td>
<td>57.2</td>
<td>70.7</td>
</tr>
</tbody>
</table>

5. Conclusions

As B2C market increases, web stores focus more on their relationship with customers. Understanding visitors’ purchase behaviors and responding appropriately in online market are important issues.

Former works on purchase prediction focus more on changes of users’ path in a website, dividing customers’ sequential orders into specific tasks, and selecting variables related to purchase. In this study, we suggest an approach based on th those customers’ sessions can be transformed into a graph, which also represented in density of a session based on graph theory. We represent the networks of visits using graph theory. The density of a network refers to the degree links among objects. Thus, high density in a network means the degree of linkage among objects is concrete, and delivery or proliferation of information of an object is dynamic and fast. Therefore, involvement or interest in products is higher and possibility of “purchase” intention increases.

Online user behaviors are differentiated into several types according to their propensities. In this study, we dichotomize “purchase” and “not purchase” group since buying or not is an important issue. Actually a small part of customers incurs a large part of total profits, that is, most of the customers are unprofitable. Using clickstream data, we model users’ sessions in web into density of a graph, which again produces DOS (Density of a Session). The performances of purchase prediction by different prediction methods including DOS are compared.
We found that all prediction methods except total visit time are useful to differentiate between "purchase" and "not purchase" group. We also conducted Logit Analysis to examine purchase prediction by each method. The results from Logit Analysis show that DOS predicts purchase behavior better compared to other predictors.

This study, however, has several limitations. Firstly, we select a limited number of products to scale down the experiment. Secondly, many sessions are deleted from original data set to have meaningful sessions which can be transformed into graphs. Thirdly, there can be more meaningful predictors to predict users purchase behavior except variables used in this study.

6. References