Tourism Marketing Plan Making Based on Data Mining Technology

Zhang Xuzhi*, Sun Lixin
Yantai Nanshan University, Longkou, 265713, China

Abstract: The popularity of the Internet and the rapid development of e-business have brought expanded development for tourism and other traditional trading industries. If the businesses develop the corresponding marketing strategies to all customers, it will not only be a waste of time and resources, but also greatly increase the cost of marketing. Therefore, businesses are also required to select some nodes with special effect to carry out marketing activities, that is to say, the core marketing group. The key to successful implementation of marketing is to accurately identify the core groups based on data mining, which can further trigger the purchasing or usage behavior of other non-core groups. This paper introduces the formulation of tourism marketing plan which is realized based on the combination of original tourism data and data mining technology. Then, it carries out analysis of the actual demand in the tourism market and recommends and formulates the relevant strategies by using data classification to find out the specific moving trends of core customer groups.

Keywords: Data mining; Marketing plan; Potential customers; Demand analysis; Cost analysis

1. Introduction

With the development of computer network technology and logistics industry, e—business based on internet technology has become an important part of the modern market. The report shows that: at the end of June, 2012, the scale of online shoppers in China had reached 210 million. Its proportion to all Internet users is as high as 39% and online shopping users have increased 15.94 million compared to the figure at the end of December 2011 with an increasing rate being up to 10.8%[1]. The market scale of e-business in China had become the largest in world in 2015[2]. In the traditional marketing strategies, the businesses usually carry out the targeted product recommendations based on consumption characteristics the users and the behavior similarity between the users. For example, according to the user's purchase history, they are able to predict the purchase behavior of users in the future. On such basis, businesses can directly market to the users and obtain more profits[3,4]. However, huge amount of evidence has shown that more and more people start to resist the traditional marketing methods and turn to believe the market information and product information obtained from the relatives or friends around them[5,6]. In this context, the concept of marketing based on data mining is developed. Marketing based on data mining is to seize the interests of customer groups and fully mobilize their enthusiasm. Based on the principle of the spread of the virus, information of product and service will be packaged into the "virus" with strong appealing so as to continuously attract the target customer groups. In this way, they will be transformed from the buyers of products or service into disseminators and recommenders of products or service. Core group plays a very important role in the marketing based on data mining and it has a series of features in the customer network which include a large number of online comment, higher click rate, more friends relationship, higher authority and higher network status. In this regards, it is able to make use of its enormous influence on the customer's purchasing will by recommending to other customers.

2. Potential Customer Mining Constructed by Network Influence Set

Currently, the research on data mining of the core group is still in the initial stage. In the literature [7], Web customer trust network is established based on the customer trust relationship. The degree of trust and the direction of trust between different customer nodes are not the same. Its expression mode is shown in equation(1):

\[ G = \{ C, E, W \} \]  

Among them, \( C \) represents a node in the network; \( E \) represents the trust relationship between nodes; \( W \) represents the trust weight on the trust relationship. In the first trust network of customers \( G, C_i (i \in [1,2,\ldots,n]) \) represents any node in the network with a lot of node information being included, such as the interest characteristics of the commodity in the client node, the recommendation of one client node to other nodes, the acceptance of recommendation from other nodes and comment score of client nodes for the goods. A core group mining algorithm based on information diffusion is proposed, which is relied on the web customer trust network. However, the complex relationship between the nodes in the web customer trust network is ignored and the construction method of the influence set for node network is not proposed[8-11].

Definition 1: network neighboring set. In the web customer trust network, a set of nodes which have direct trust relationship with one node is called the neighboring set of the node. Hereby, we use \( \text{neighbour}_{C_i} \) to present the neighboring set of \( C_i \). The set has included all the nodes: the trust node \( C_i \) and the nodes being trusted by \( C_i \). The specific expression form is shown in equation(2):

\[ \text{neighbour}_{C_i} = \{ \text{neighbour}_{C_{i\text{parent}}}, \text{neighbour}_{C_{i\text{child}}}, \ldots \} \]  

Among them, \( \text{neighbour}_{C_{i\text{parent}}} \) represents the set of nodes which are trusted by \( C_i \). And it is called the parent network neighboring set of the node \( C_i \). The \( \text{neighbour}_{C_{i\text{child}}} \) represents the set of nodes which are trusted by \( C_i \). And it is called the child network neighboring set of the node \( C_i \).

The node \( C_i \) has direct trust relationship with the nodes in the network neighboring nodes. But the trust weight \( W \) can only represent the degree of trust between nodes, but it can not accurately represent the influence of the trust relationship of nodes on the purchasing behavior, therefore, we introduce the concept of influence degree.

Definition 2: Node recommendation position. Node recommendation position a node represents the position and role of node in the recommendation process. The calculation formula of the width of node recommendation is proposed in equation (3):
Among them, \( B^X_{C \rightarrow C_j} \) represents the recommendation position width of commodity of X category made by \( C_i \) to \( C_j \). \( m_{C \rightarrow C_j} \) represents the recommendation times made by \( C_i \) to \( C_j \). \( m_{C \rightarrow C_j} \) represents the recommendation times made by \( C_i \) to \( C_j \) in the neighboring set of the child-network. If \( m_{C \rightarrow C_j} \) represents the recommendation times made by \( C_i \) to \( C_j \) in the neighboring set of the parent-network. The smaller the \( B^X_{C \rightarrow C_j} \) is, the recommendation effect made by \( C_i \) to \( C_j \) will be larger, vice versa.

**Definition 3: Influence degree.** Influence degree is the measurement of influence degree of customer recommendation behavior based on the Web customer trust network on the purchase behavior. It is represented by \( E \). The influence degree \( E \) is a directed value, its direction is in the opposite of that of the trust relationship \( W \).

The influence degree \( E \) will carry out calculation based on the trust degree, the user's comment on the goods, the recommendation between nodes and other factors. Moreover, it is also processed by normalization and the specific calculation formula is shown in equation (4):

\[
E^X_{C \rightarrow C_j} = \frac{W_{C \rightarrow C_j} \times (R_{ij} - \overline{R})}{\sqrt{\sigma^2_{R}}}
\]

Among them, \( E^X_{C \rightarrow C_j} \) represents the influence of \( C_i \) on \( C_j \). \( W_{C \rightarrow C_j} \) represents the trust degree of \( C_i \) to \( C_j \). \( B^X_{C \rightarrow C_j} \) represents the width of recommendation position of \( C_i \) to \( C_j \) about commodity X. \( R_{ij} \) represents the comment score of \( C_i \) for the commodity. \( \overline{R} \) represents the average comment score of customers for the commodity. \( \sigma^2_{R} \) represents the standard deviation of the commodity score of \( C_i \). Among all customers. Influence degree includes the degree of trust between the customer nodes, the commodity score of customer nodes and the recommendation effect between nodes, which can reflect the influence degree of the client node on the other nodes in terms of the purchasing behavior.

**Definition 4: Skewness of average influence degree.** The deviation of influence degree of nodes in the parent network on the node relatively to the average influence degree is called the Skewness of average influence degree, as shown in equation (5).

\[
\sigma_{1 \rightarrow 2 \rightarrow N \rightarrow}^{C_{influence}} = \left| \sum_{C_{neighbour}} \frac{E_{C \rightarrow C_j}}{E_{C_j}} - \overline{E_{C_j}} \right|
\]

Among them, \( \sigma_{1 \rightarrow 2 \rightarrow N \rightarrow}^{C_{influence}} \) represents the skewness of average degree influence of node \( C_i \) to node \( C_j \). \( \overline{E_{C_j}} \) represents the average value of influence degree for nodes in the neighboring set of parent network of node \( C_j \).

The calculation formula of threshold value \( K \) is shown in equation (6):

\[
K = \frac{\sum_{C_{neighbour}} \frac{E_{C \rightarrow C_j}}{E_{C_j}} \cdot \overline{E_{C_j}}}{m_{C \rightarrow C_j}}
\]

The influence degree of the neighboring set in the parent network is less than the node of average influence degree. If the skewness of average influence degree is greater than \( K \), it will be removed from the neighboring set in the parent network of the node.

Construction algorithm BUNIS of network influence set is shown in the following content:

1. The network neighboring set Neighbor is extracted from Web customer trust network \( G \);
2. Obtain the node information which mainly includes the recommendation times to other nodes, the recommendation times received and accepted from other nodes. Then, it will calculate the width of recommendation position of the node;
3. Based on the trust degree between nodes, the width of recommendation position, the comment score of the customers, the influence degree \( E \) between nodes is calculated.
4. The threshold value \( K \) is calculated and then the value is used to deal with the relationship between the nodes. Then the network influence set Influence is obtained.

The detailed description for the construction algorithm of the network influence set is shown in the following content:

Input: customer trust network \( G \), the recommendation times of the node \( \left( m_{C_1, \text{in}}, m_{C_2, \text{in}}, \ldots, m_{C_X, \text{in}} \right) \), the recommendation times received by the node \( \left( m_{C_1, \text{out}}, m_{C_2, \text{out}}, \ldots, m_{C_X, \text{out}} \right) \), the comment vector of customers \( \left( R_1, R_2, \ldots, R_N \right) \).

The algorithm BUNIS can be used get the network influence set of nodes and its specific form is shown in equation (7).

\[
influence_{C} = \{ influence_{C \rightarrow \text{parent}} \cup influence_{C \rightarrow \text{child}} \}
\]

Among them, \( influence_{C} \) represents the network influence set \( C \), \( influence_{C \rightarrow \text{parent}} \) represents the set of nodes which will influence node \( C \). And it is called the parent network influence set of node \( C \). \( influence_{C \rightarrow \text{child}} \) represents the set of nodes influenced by node \( C \). And it is called the child network influence set of node \( C \). The network influence set of nodes is an important basis for core group mining.
of marketing. The network influence set constructed in this paper can reflect the complex relationship between the nodes in the actual Web customer trust network and it is also of great significance for the subsequent experimental work.

3. Marketing Plan Model for Group Recommendation

In this paper, the recommendation model of core group based on the network influence set is shown in Figure 1. The model fully considers the exerting model of the core group and the recommendation history core node is used as an important basis for the recommendation. We match the characteristics of the goods in accordance with the historical recommendation of core nodes and the individual interests of the core nodes to calculate the degree of recommendation.

Data pre-processing

Extract the characteristics of commodities to be recommended

Calculate the matching degree between commodities to be recommended and commodities characteristics of the nodes in historical purchase

Extract the commodities characteristics of core nodes in network influence set historical purchase

Extract the commodities characteristics of core nodes in network influence set historical purchase

Calculate the degree of recommendation

Recommendation list

Figure 1. Core Group Recommendation Model Based on network Influence Set

In order to accurately describe the core group recommendation model, we give the following relevant definitions:

Let \( n \) characteristics are abstracted from the commodities of certain categories. The \( n \) th characteristic is corresponding to \( m_i (i \in \{1,2,...,n\}) \) attribute values, therefore, there are \( \prod_{i=1}^{n} m_i \) commodities with different characteristic combination. We use \( X_k \) to represent the commodity of \( k \) th characteristic combination, among them, \( k \in \{1,2,...,\prod_{i=1}^{n} m_i\} \).

Definition 5: matching degree of characteristics. It indicates the degree of the similarity between the characteristics of a commodity and the interest characteristics of customer for this category of commodity.

\[
M = \sum_{i=1}^{n} \sum_{u=1}^{m_i} L_{iu} \tag{8}
\]

Among them, \( m_i \) represents the number of characteristic value of \( i \) th characteristic, \( L_{iu} \) represents the mark of characteristic value which is 1 or 0.

Definition 6: recommendation degree. It represents the degree for a commodity with certain characteristics worthy of being recommended to a customer node. We use \( P_{C_j,X_k} \) to represent the recommendation degree of commodity of X category with \( k \) th characteristic to node \( C_j \). The bigger the recommendation degree is, the worthier the commodity with such characteristic being recommending to the customer node is. The specific calculation method is shown in equation (9):

\[
P_{C_j,X_k} = \frac{m_{c_j,self} \times M_{c_j,self} + m_{c_j,recom} \times M_{c_j,recom}}{\sqrt{m_{c_j,self}^2 + m_{c_j,recom}^2}} \tag{9}
\]

Among them, \( k \in \{1,2,...,\prod_{i=1}^{n} m_i\} \), \( m_{c_j,self} \) represents the number of commodities with \( k \) th characteristic combination bought by node \( C_j \), \( m_{c_j,recom} \) represents the number of commodities with \( k \) th characteristic combination recommended by node for the child network influence set. \( \prod_{i=1}^{n} m_i \) represents the number of characteristic combination of commodity. \( M_{c_j,self} \) represents the characteristic matching degree of commodity bought by node \( C_j \). \( M_{c_j,recom} \) represents the characteristic matching degree of commodity bought by child-network influence set of node \( C_j \). Its calculation equation is shown equation (10):

\[
M_{c_j,recom} = \sum_{c_i \in \text{influence}_{C_j,child}} \left[ \frac{m_{c_i,self} \times M_{c_i,self}}{m_{\text{influence}_{C_j,child}}} \right] \tag{10}
\]
In the calculation of the recommendation degree, we not only consider the interest degree of the core node itself, but also consider the interest degree of node to the commodity in the child network influence set. Therefore, it can describe the status and function of the core node in the Web customer trust network.

4. Experimental analysis

4.1. Experimental preparation

The hardware and software conditions of the experimental will affect the results to a large extent. The hardware and software environment of this paper are shown in Table 1.

The experimental data uses the real data collected from Epinions website, including 75887 nodes (received recommendation time, recommendation times being sent out, comment scores), 508840 trust relationship (trust relationship and trust weight).

Table 1 Experimental Configuration

<table>
<thead>
<tr>
<th>Name of configuration</th>
<th>configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Intel(R)Core(TM)i5-460M2.53GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>2G</td>
</tr>
<tr>
<td>Hard disk</td>
<td>500G</td>
</tr>
<tr>
<td>Operating system</td>
<td>Windows7 flagship version</td>
</tr>
<tr>
<td>DBMS</td>
<td>Mysql5.5</td>
</tr>
<tr>
<td>Development language</td>
<td>JAVA(JDK1.7)</td>
</tr>
</tbody>
</table>

In order to analyze the content of the study in a better way, this paper will carry out experiments based on the three data sets shown in Table 2. According to the construction principle of the network influence set, the network influence set of nodes is constructed and the core group excavation algorithm based on the network impact set is used to excavate the core group.

Table 2 Experimental Data Set

<table>
<thead>
<tr>
<th>Number of nodes</th>
<th>DataSet1</th>
<th>DataSet2</th>
<th>DataSet3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust relationship</td>
<td>2867</td>
<td>5364</td>
<td>7599</td>
</tr>
</tbody>
</table>

4.2. Analysis of Network-Coverage

Network-Coverage (NC) refers to the ratio of the nodes receiving the commodity information in to all network nodes in the customer trust network. It is shown in equation (11).

\[
NC = \frac{m}{M}
\]  

(11)

Among them, \(m\) represents the number of nodes of commodity information that is received in the customer trust network, \(M\) represents the number of all the nodes in customer trust network.

![Network-Coverage vs Number of core nodes](image)
This paper carries out experiment and analysis of the influence of the number of core nodes on the influence of network influence node based on three experimental data sets. As shown in figure 2 (a) and (b), it is seen that with the increasing of the number and proportion of core nodes in the customer network, the growth area for network coverage rate is stable and when the ratio of core node is very small, it can still achieve good network coverage. The network coverage can be a reflection of the promotion of commodity information. When the businesses carry out marketing based on data mining, they hope that the commodity information of them is received by more customers or potential customers. Therefore, the network coverage of the core node is a major aspect concerned by businesses.

4.3. Accuracy Analysis

In this paper, the average absolute error ratio MAE of sales is used as the evaluation criterion for recommendation accuracy of the core group[9]. Its definition is:

\[ \text{MAE} = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{\text{Count}_{\text{recommend}}(i) - \text{Count}_{\text{real}}(i)}{m \times \text{Count}_{\text{real}}(i)} \right) \]  

Among them, \( m \) represents the number of core nodes, \( \text{Count}_{\text{recommend}}(i) \) represents the estimated sales recommended to core nodes, \( \text{Count}_{\text{real}}(i) \) represents the actual sales. The smaller the MAE is, the more accuracy of the system will be, vice versa.

In order to verify the accuracy of the recommendation to the core group, this paper carries out the experiment in three data sets respectively. Firstly, the proposed algorithm can be used to calculate the value of the recommendation degree. In fact, it can be realized by forecasting the purchase quantity. The average error ratio of the estimated value and the actual value is calculated by the formula (12). As shown in Figure 3, the horizontal axis represents the number of core nodes, the vertical axis represents the average absolute error ratio. It can be seen that the model recommended to the core group proposed in this paper has a better accuracy. At the same time, with the increasing of core group, the recommendation model can still maintain a good stability with the increasing of the core group.

5. Concluding Remarks

Presently, the research on the core group recommendation is just started. The research work in the future includes the optimization of recommendation model and the recommendation algorithm, the formulation of marketing strategy of the core group and so on. Marketing based on data mining is a kind of advanced marketing strategy which is closely combined with the modern technology and there are a lot of problems to be solved in this regard. Among them, the perfect marketing mechanism based on data mining is the necessary condition for the successful implementation of marketing based on data mining. At present, the research of marketing based on data mining is still relatively scattered, therefore, it is recommended to establish a complete marketing mechanism based on data mining.

References


