Research on Tourism E-commerce based on Data Mining

Yan Liu¹,²,a
¹School of Economics and Management, Southwest JiaoTong University, China
²College of Tourism, Sichuan Agricultural University, China
ayanliu0422@sina.com

Abstract—This paper describes in detail the web data mining technology, analyzes the relationship between the data on the web site to the tourism electronic commerce (including the server log, tourism commodity database, user database, the shopping cart), access to relevant user preference information for tourism commodity. Based on these models, the paper presents recommended strategies for the site registered users, and has had the corresponding formulas for calculating the current user of certain items recommended values and the corresponding recommendation algorithm, and the system can get a recommendation for user.

Keywords-data mining; tourism e-commerce; web data mining; recommended system.

I. INTRODUCTION

Tourism electronic commerce development main have portal site, professional tourism website, the traditional tourism business travel website at present. The content of these sites is mostly simple introduction of enterprises, the main tourist routes, attractions, go out to try at home and travel works etc.. The technology is only static or dynamic not static website Webpage custom style, no way to achieve rapid update framework. There are also problems in system porting, cannot cross platform, scalability is poor, no component construction thought. But China's tourism electronic commerce transactions are still in the traditional mode: (1) the hotel reservations in the phone book, the payment; (2) to telephone booking, ticket booking door payment; line booking by telephone booking, payment etc.. [2]; (3) travel website should establish tourism information perfect system, and component extension, can be updated framework, technology should be platform for developing multi-tier architecture model.

In many of these tourism e-commerce and frequent trading, people have made quite a lot of data, this to other people or business institutions researchers numerous potential opportunities to gain insight into the user experience, business marketing, personal preferences and usually so-called human behavior [1]. In this paper the method of using Web data mining, can design a recommended offer for tourism e-commerce system.

II. ARCHITECTURE OF TOURISM E-COMMERCE PLATFORM BASED ON INFORMATION RELEASE SYSTEM

Tourism e-commerce or e-government framework (Fig.1) contains the presentation layer, application layer and application support layer, system software support layer, physical layer network layer. Based on the general framework of e-commerce platform information release system (Fig.2) belongs to the e-commerce or e-government framework of the application layer and the application support layer, that can build up in the system software support layer of this framework, provides the application layer and the component services application support layer to layer.

The platform is divided into 2 levels: (1) electronic commerce platform on the information release system; (2) information publishing system based on the characteristics of electronic commerce website. Based on the platform of electronic business information publishing system, including platform supporting collaboration system, platform customized website mode code database (data layer, on the platform outside the system belongs to the platform but part of a system), system, e-commerce platform for information construction of e-commerce platform release system, platform template parsing system and platform system.

Based on the information system of the characteristics of the electronic commerce site features, including hotel reservations, car rental, meeting scheduled assembly components, assembly line scheduled predetermined components and ticket booking module etc.. These components formed the e-commerce portal on real significance present for customer. User interaction with the platform of business process is divided into the following two categories:

(1) Process of electronic commerce platform
Platform to build website construction platform framework through system memory from the model code and call mode code, also call platform supporting collaboration system, complete with their business outside support function. In addition, platform information system

Figure 1. The overall architecture of e-commerce platform
according to the model building system requirements part
code loading platform, construct information publishing
framework, also call platform supporting collaboration
system, complete with their business outside support
function.

Figure 2. The general framework of e-commerce platform information
release system

Platform template parsing system release system and
platform from the platform construction of information
system are extracted from the filled XML data tree
generated framework and model code data, and filled with
data XML file for conversion of XSLT template parsing,
finally generate the HTML file. Platform system
extraction to generate the HTML file from the platform
template parsing system are combined, show portal
components to generate custom.

(2) The user initiated process
Website login users fill out web forms for information
exchange, information from the site into the system, after
treatment with the underlying data layer interaction, and
then along the opposite direction results, show the user
finally.

III. RECOMMENDED STRATEGIES FOR TOURISM E-
COMMERCE

(1) Items similarity
Item similarity refers to the universal existing
relationship of the item i and item j when customer buy it,
with sim (i, j) expression. Corresponding to the actual
business operation, sim (i, j) said to buy things on the site
of the user, if they buy the goods “i” will usually buy
goods “probability. In other words, sim (i, j) represent
probability of i and j items is purchased together. The
greater the probability, we say that the two items more
similar. This model can be used to guide the
implementation of cross selling task recommendation
system.

Calculate similarity based on cosine: using this
method, we put the two item i and j as two dimensional
vector m user space. The computation of similarity is
calculated the angle between the vectors representing
these two items of cosine. If the user “u” purchased items
i, then matrix in section u for the “i” column value is 1,
otherwise 0. In general, i and j similarity calculation
formula for goods is:

\[
sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{|i| \times |j|}
\]

\( \vec{i} \cdot \vec{j} \) Indicates the number of product of vector, |i|
and |j| is respectively norm of vector.

(2) The user clustering model
According to the user visit on the site to generate
clustering, similar access behavior of users will be
clustered into a class. It is worth noting that for the
registered users, each registered users themselves to form
a category.

Transaction clustering, using [3] the most widely used
K-means algorithm in data mining. Before using the
clustering algorithm K-means transaction, the transaction
must be expressed as a numeric vector K-means to
operation. In order to vector to conduct the affairs of
representation, we can put the transaction mapping in the
page space, a transaction is expressed as a multi-
dimensional vector space. Set up site all web pages for
P={p1, p 2,..., pn}, the transaction T can be
expressed t=(w(P1, t), w(P2, t),..., w (Pn, t) in multi
dimension vector page space P, which:

\[
w_{pj} = \begin{cases} 
\gamma \text{ number of page p that appear in affair } t \\
0 \text{ number of page p that don't appear in affair } t
\end{cases}
\]

Identified transactions are transformed into a
multidimensional vector on the space pages P, the uid is
the clustering algorithm based on K-means for empty the
affairs.

(3) Establish items - Page Mode
We need to analysis and identify the page of the goods
involved information extraction methods of data mining.
Information extraction methods on the site on the HTML
or XML page uses information extraction, extract the
relevant items of information from HTML or XML, you
can use [2,20] method, the use of machine learning theory
to the items of information extraction.

The information extraction of HTML and XML can
obtain the weights W (I, P) that relative to this page or
each page relates to the items set and these items, for
example, the weight can be defined as frequency of the
goods i appear on the pages P. Reset all items web site
involves the composition space for the Is (Item Space),
the space size is n. So in the information extraction, each
page can be represented as an n-dimensional vector
objects in P space, establishing item - Page
Mode p = \{w_{i1}, w_{i2},..., w_{in}\} . Among them:

\[
w_{ip} = \begin{cases} 
\frac{w(i, p)}{weight of goods i in page p} \text{ if page p that related to goods i } \\
0 \text{ otherwise } \end{cases}
\]

1≤i≤n

670
(4) The establishment of user clustering model about items

Through the web usage mining produces a user clustering similar access behavior, information extraction through web content mining produced a page in the object space Is vector representation. Using them as the basis, can produce goods - user clustering model. Item - user mode is a kind of user to goods space Is on an item's interest, denoted as: $\phi(i, c)$ or $\phi(i, u)$. The formula for items - user clustering model:

$$\phi(i, c) = \frac{1}{|k|} \sum_{i=1}^{n} w(i, p) \times \text{weight}(p, c)$$

$$\phi(i, u) = \frac{1}{|u|} \sum_{j=1}^{n} w(i, p) \times \text{weight}(p, u)$$

(4)

Where $w(i, p)$ is weight of items i relative to the page p, weight $(p, c)$ or weight $(p, u)$ is p relative to c or u page clustering weights. N is a space the size Is of the items. Then for every cluster c or u can be expressed with respect to the n-dimensional vector the clustering weights for space objects, user clustering model items:

$c = \langle \phi(1, c), \phi(2, c), ..., \phi(n, c) \rangle$ or $u = \langle \phi(1, u), \phi(2, u), ..., \phi(n, u) \rangle$

(5) Recommendation strategy

For the recommendation of the registered users, we mainly consider 4 main factors from strategy.

1) Evaluation value $\text{Ru}$ of item i that ever buy by user;
2) Correlation similarity $\text{Sim}(i, j)$ between users have bought items i and recommended values are calculated items;
3) The date and time DT(i) of item i that was choose;
4) Item-user clustering mode $\phi(j, u)$ of Items j that the recommended values are calculated when the user access to the website long time ago;

The calculation formula of goods j that the goods whether goods j can be recommended to the user u are as follows:

$\text{Rec}_u(j) = \phi(j, u) + \sum_{i=1}^{n} \text{Sim}(i, j) \times (R_u + 1) \times T(i)$ (5)

“i” is a purchased items of user u during a period of registration in the database, or is the items in the shopping cart.

Parameter $R_u$ is the evaluation value that the user have bought items. Here use $R_u + 1$ as a multiplication factor, in order to avoid is equal to 0, the items in the formula in the other involved in the calculation of the value of failure.

$\text{Sim}(i, j)$ is associated similarity that calculated articles i and j earlier. Used as a factor in this formula, said that if the correlation similarity of user u that have high values for purchased item i and item j, it has also a great possibility that users buy goods i and j. So $\text{Sim}(i, j)$ affect the calculation recommended value items j in the very great degree.

Parameter $\phi(j, u)$ is a reflection of the user u used to visit the item j. Here we consider the user u has high value access to the goods j, the more he is interested in the good j.

$T(i)$ is the item I purchase date time and the computation time of a function of two variables: $T(i) = f(DT(), DT(i))$. If the purchase date and time for good i from the current date is long, so it can produce recommended value is small, or more. So $T(i)$ in the formula is a decreasing function of a mathematics.

IV. THE OPERATING STEPS OF RECOMMENDATION ENGINE

System find similarity relation according to the relationship between the tourism goods and articles involved they request the latest page, find can recommend items. If an item in the current user requests the page weight high, so high and the articles of association of the similar items can be used as recommended object.

When users browse the site, the session manager (SM) responsible for communicate with the user, to capture the user at the same time related information and send it to the dialogue manager (DM), to call in maintaining and processing context and user interaction and recommender (RE), in the recommenders to complete the calculation recommend the result set, the set of session manager information, the sender shall submit the results to the Web server, complete the recommendation work. Among them, the context of user interaction with the system is recorded in the dialogue context.

V. CONCLUSIONS

As we realize the separation of offline part and online part, which can adapt to the large amount of data, greatly improve the scalability of personalization recommendation service and real-time response speed. At the same time, the integration of Web mining technology, in the use of data is relatively small, or Web site contents change frequently, also can provide high quality and personalized recommendation service. Through the interaction of the two module, this system can guarantee the finally presented to customers recommend collection in real time, the new, it is reasonable.

REFERENCES