ASSOCIATION RULE ANALYSIS FOR TOUR ROUTE RECOMMENDATION AND APPLICATION TO WCTSNOP

FANG Hui[a,b], CHEN Chongcheng[b,*], LIN Jiaxiang[c], LIU Xianfeng[d], FANG Dong[d]

[a] Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Spatial Information Research Centre of Fujian, Fuzhou University, Fuzhou University District Xueyuan Road, China-chencc@fzu.edu.cn
[b] Dept. of Computer Science, Fujian Provincial Key Laboratory of information Processing and Intelligent Control, Minjiang University, Fuzhou University District Xiyuangong Road, China-fh_fzu@126.com
[c] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, China
[d] Fuzhou Silviscene Information Technology Co. Ltd, China

Commission IV, WG IV/3

KEY WORDS: Association Rule Analysis, FP-growth, Cultural Tourism Service, E-tourism, Route Recommendation

ABSTRACT:

The increasing E-tourism systems provide intelligent tour recommendation for tourists. In this sense, recommender system can make personalized suggestions and provide satisfied information associated with their tour cycle. Data mining is a proper tool that extracting potential information from large database for making strategic decisions. In the study, association rule analysis based on FP-growth algorithm is applied to find the association relationship among scenic spots in different cities as tour route recommendation. In order to figure out valuable rules, Kulczynski interestingness measure is adopted and imbalance ratio is computed. The proposed scheme was evaluated on Wangluzhe cultural tourism service network operation platform (WCTSNOP), where it could verify that it is able to quick recommend tour route and to rapidly enhance the recommendation quality.

1. INTRODUCTION

The fast-growing, tremendous amount of data in E-Tourism, which are collected and stored in large and numerous data repositories, have far exceeded users' ability for catching useful information by themselves. Users are eager to know where is popular, where to go first and where to travel next by the way, which is known as Tour Route Planning, especially when browsing E-tourism websites. However, the traditional E-tourism websites offer only the query for hot route list by number of days or Destination. The users would be caught in a route data rich but information poor situation. Therefore, by identifying the characteristics of different users' needs, information recommendation would solve this problem very well. Recommender systems (RS) is first proposed to recommend for users according to their taste (Resnick et al., 1994).A comprehensive understanding that recommendation can be given in the background of data mining is elaborated(Ricci et al., 2010). This probably due to that the existing RS is obviously not competent to process data in the speed aspect, heterogeneous data aspect or data missing aspect. Therefore, how to find an appropriate way to enable more quick recommendation seems to be especially important. What's more, the widening gap between data and information calls for data mining tools that can turn data tombs into "golden nuggets" of knowledge. Data mining is the process of extracting valid and maybe unknown information from large database and then utilizing information to make crucial business and strategic decisions. To be specific, sampling techniques and dimensionality reduction techniques can be applied in the pre-processing step; classification method can be used to derive a model-based RS or content-based RS; Clustering algorithms is used to improve performance of RS; Association rules offer an intuitive framework for recommending items whenever there is a transaction. Association rule is first applied for recommender system (Fu et al., 2000). The user's future chose is predicted on his/her past experience. Then the items are listed for him/her with some support. In the field of E-tourism, huge amounts of transaction data bring a straightforward opportunity to obtain useful information through data mining (Liao et al., 2010; Lucas et al., 2013; Li et al., 2015).

2. LITERATURE REVIEW

Data mining refers to knowledge mining from large amount of data. Many people view data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others treat data mining as merely an essential step in the process of knowledge discovery. Actually industry or in the research milieu, the term "data mining" is often used to refer to the entire knowledge discovery process. Therefore, by integrating the perspectives, a broad view of data mining is adopted. i.e., Data mining is the process of discovering valuable patterns and knowledge from large amount of data. The data source can cover databases, data warehouses, Web information or other repositories.

* Corresponding author

2.1 Association Rule Analysis

Association Rule Analysis is one of the most active directions in data mining research. As the name implies, it's about seeking associations between different features in datasets. It's an important branch of data mining. Let $I = \{I_1, I_2, ..., I_n\}$ be an itemsets. Let D be a set of database transaction where each transaction T is the subset of I .Assume SA to be a nonempty set of items, a transaction T would contain SA if $SA \subseteq T$ .An association rule is a hint of the form $SA \Rightarrow SB$,where $SA \subseteq T$, $SB \subseteq T$, $SA \bigcap SB \neq 0$ .The rule $SA \Rightarrow SB$ takes effect in the transaction set D with support s ,where s is the percentage of transactions in D that contain $SA \bigcap SB$, and with confidence c ,where c is the percentage of transactions in D containing SA that also contain SB .They represent the correctness and importance of the rule respectively. The confidence of rule $SA \Rightarrow SB$ can be easily derived from the support count of SA and $SA \bigcup SB$ .Thus, the problem of mining association rules can be reduced to that of mining frequent itemsets, which will occur at least as frequently as a predetermined minimum support count.

2.2 Efficiency of Association Rule Analysis

As a basic algorithm of finding frequent itemsets for Boolean association rules, Apriori algorithm (Agrawal et al., 1993; Agrawal et al., 1994) employs approach known as a level-wise search, where k-itemsets are used to generate (k+1)-itemsets. When there is a large number of frequent itemsets, Apriori algorithm will repeatedly scan the database. To improve the efficiency of the level-wise generation of frequent itemsets, many variations of Apriori algorithm have been proposed. For instance, Hash-based technique can be used to reduce the size of the candidate k-itemsets by hashing itemsets into corresponding buckets. Partitioning technique requires just two database scans to mine the frequent itemsets through partitioning the data to find candidate itemsets. Sampling technique mines on a subset of the given data at the cost of some degree of accuracy .However, these techniques still suffer from a huge number of candidate itemsets and repeatedly scanning the whole database. Thereby, an algorithm named FP-growth (Han et al., 2000), which transforms the problem of finding long frequent patterns into searching for shorter ones in much smaller conditional databases recursively. It has been proved to be efficient and scalable for mining both long short frequent patterns, and is about an order of magnitude faster than the traditional Apriori algorithm.

2.3 Quality of Association Rule Analysis

Association rule analysis algorithm is the core of the recommender system, and the recommendation effectiveness depends on the quality of the rules. With the increase scale of database, the quantity of association rules containing redundant rules is explosively increasing. Whether a user receive a valuable recommendation from the recommender system based on association rule have been discussed further (Han et al., 2006).It would be the most favorable to infer more information from fewer rules. i.e., association rule $SA \Rightarrow \{SB, SC\}$ provides more important information than $SA \Rightarrow SB$. Several literatures have pay attention to the rules quantity problem and proposed simple association rules (chen et al., 2002), optimal association rules (Li et al., 2002) and minimal association rules (Bastide et al., 2000). However, these forms of rules are not fit for applications of personalized recommendation, which concerns the rule quality instead of quantity. Ding come up with an idea that the number of association rules could be cut down with mergence of different rules (Ding et al., 2015). Wang proposed a kind of association rule mining algorithm with maximal nonblank for personalized recommendation (Wang et al., 2004). Specially, several association rule mining algorithms based on interestingness are presented to solve the existence of uninteresting or useless rules (Geng et al., 2006).Of these, more attention are given to objective interestingness measure. Lift measure takes into account the support of SB in order to filter the negative rule (Hussein,2015). But whether the confidence is proportional to the lift is problematic. The interestingness model based on difference in thought is mainly to introduce negative items to improve the validity of association rules and it skillfully combined with the support of the rule and its confidence to deal with the relationship between the two thresholds (Zhou et al., 2000). However, the asymmetry of the positive and negative calculation exists. Piatetsky-shapir(PS) measure exploits the orientation of correlation between SA and SB based on the probabilistic theory (Piatetsky-Shapiro,1991). Yet, its prerequisite is that mutual interestingness measures are the same. Unfortunately, most of the measures above do not have the null-invariance property. Because large data sets typically have many null-transactions, it is important to consider the null invariance property when selecting appropriate interestingness measures for pattern evaluation. Different from them, Kulczynski(Kulc) measure with zero invariant properties can deal with null-transaction itemsets, which exclude any investigative object(Han et al., 2006).

On the whole, in order to improve efficiency and quality of association rule analysis together, FP-growth algorithm is adopted as data mining method and Kulc measure is recommended for rule pattern evaluation.

3. DESCRIPTION

3.1 Model Description

The project is implemented as a sub-application of WCTSNOP(Silviscene Information Technology Co. Ltd,2016). As Fig.1 shown below, the model of route recommendation service is made up of user interface module, user profile database, data mining module route database, POI database, and rule database.When a user logins the user interface, the user interface module refers to user profile database to get profile detail as the reference of route recommendation. Then the user can browse and query the route arrangement. Data mining module applies Apriori algorithm or improved Apriori Algorithm to find out frequent itemsets. For purpose of finding frequent itemsets, the data mining module scans the route database circularly. The association rules generated by data mining module are stored in rule database. Those rules which satisfy minimum support value and minimum confidence value are filtered by rule evaluator. These POI IDs in an examined rule are offered to the POI database and complete details of POI IDs are retrieved as recommended POIs. The routes which consist of the recommended POIs are provided to the user. The POIs may be a city or a scenic spot.

3.2 Experiment Description

The experiment is tested on WCTSNOP, which is a comprehensive O2O E-tourism information service platform. It is developed and managed by Fuzhou Silviscene Information Technology Co. Ltd. The platform provides tourist location-based information service, virtual experience service, tour route planning service, virtual community service and other functional services. In the experiment, computer configuration is Inter(R) Core(TM) i5-3470 CPU @3.20GHz,internal storage 8.00GB and hard-disk space 800GB. SQL2008 R2 serves as the database Visual Studio 2010 is employed as development tool. The experimental data are selected from 1244 registered users on Wangluzhe website. A total of 1244 users are randomly distributed at all ages. The sex ratio is about 1:l and jobs and income levels also follow the principles of diversity. In route databases, route data include 1337 transactions from ten provinces: Fujian, Guangdong, Zhejiang, Jiangxi, Guangxi, Jiangsu, Hubei, Shanghai, Shanxi, and Xinjiang.

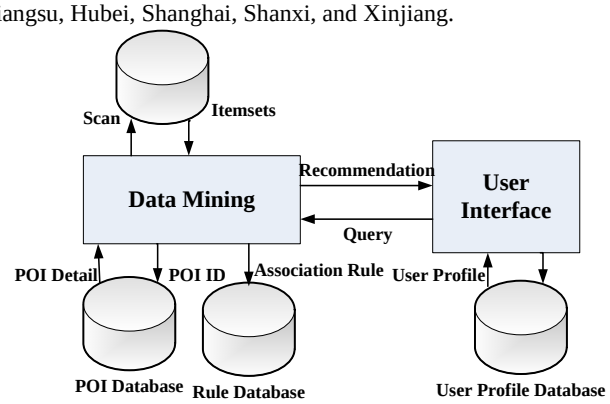

Figure 1. Route recommendation service model

3.3 Data Description

The route data in route database includes total route data and single day route data. A total route data contains several single day route data. Their data structure are presented in Table I and Table II below (where num=0, 1, 2, ...). A route includes several different cities and a city includes several different scenic spots.

Symbol	Description	Value
ID	Total route identification	tr_num
Name	Total route name	XX city X days tour
StartID	departure city identification	c_num
DestID	Destination city identification	c_num
Day	Total number of days	num
UserID	User identification	u_num

Table 1. Total route data

Symbol	Description	Value
ID	Total route identification	sr_num
RoutID	Attached total route identification	tr_num
CityID	Current city identification	c_num
SpotID	All scenic spots identification in single day	sp_num-sp_num-sp_num…
Dayord	Day order	num

Table 2. Single day route data

4. IMPLEMENTATION

4.1 Data Preprocess

Data Preprocess are constituted by three parts: data clean, data selection and data transformation. At first step, outliers are modified, missing values are deleted, and multiple data sources are combined. The resulting data are stored in a data warehouse. In detail, abnormal single day IDs are figure out through comparing the total route data and the single day route data. Correct single day IDs and merge two data tables into one data table according to the same RoutID. The route transactions are reduced to 1298. The merged data structure is shown in Table III.

Symbol	Description	Value
ID	Total route identification	mr_num
RoutID	Attached merged route identification	tr_num
StartID	Departure city identification	c_num
CityID	All cities identification in merged route	c_num-c_num-c_num…
SpotID	All scenic spots identification in merged route	sp_num-sp_num-sp_num…
Day	Total number of days	num
Dayord	Day order	num

Table 3. Merged route data

4.2 Data Mining

Select FP-growth algorithm (Han et al., 2000) as Data mining method. The first scan of the merged route database is the same as Apriori, which derives the frequent 1-itemsets and their support counts. Let the minimum support count to 2.The set of frequent items is sorted in the order of descending support count. Then an FP-Tree is constructed as follows. First, the root of the tree is created with “null” label. Scan database D a second time. The Scenic spots in each route are sorted according to descending support count, and a branch is created for each route. When another branch share a common prefix, increase the count of the node by 1.Then create a new node, which is linked as a child to the prefix. To facilitate tree traversal, a scenic spot header table is built in order that each scenic spot points to its occurrences in the tree via a chain of node-links. In this way, the problem of mining frequent patterns in databases is transformed into that of mining the FP-tree. The FP-tree is mined as follows. Start from each frequent length-1 pattern. Construct its conditional pattern base, which consist of the set of prefix paths in the FP-tree co-occurring with the suffix pattern. Next, construct the corresponding conditional FP-tree. Perform mining recursively on the tree. The pattern growth is achieved by the joint of the suffix pattern with frequent pattern generated from a conditional FP-tree. In view of huge amount of merged route database, FP-tree is

constructed to projected database, which is divided from the raw database.

The pseudo-code of FP-growth algorithm is given as:

procedure FP_growth*(Tree,a)*

if *Tree contains a single path P* ***then***

for each *node in path P merge one by one, denoted as b*

generate pattern $a \cup b$ *with support_count=minimum support_count of nodes in b;*

else for each a_i *in the header of Tree{*

generate pattern $b = a_i \cup a$ *with support_count=* $a_i \cdot$ *support_count;*

construct b *'s conditional pattern based and then* b *'s conditional FP_tree Tree* $_b$ *;*

if *Tree* $_b \neq \phi$ ***then***

call FP_growth(Tree $_b$ *,b);***}**

4.3 Pattern Evaluation

Although the minimum support value and minimum confidence value help to filter the majority of uninteresting rules, nonsense rules still exist. After FP-growth mining, the truly interesting pattern representing knowledge based on interestingness measures should be identified. The Kulc measure with Imbalance Ratio is adopted due to its soundness and insensitivity. The Kulc measure of SA and SB is to average two confidence value, as follow:

$$Kulc(SA,SB) = \frac{1}{2}(P(SA|SB) + P(SB|SA)) \quad (1)$$

i.e., average the probability of scenic spot set SA given scenic spot set SB and the probability of scenic spot set SB given scenic spot set SA .If the resulting value is greater than the given threshold, then SA and SB are meaningfully correlated, meaning that the occurrence of one implies the occurrence of the other. Otherwise, SA and SB are meaninglessly correlated, meaning that the occurrence of one leads to the absence of the other. Next, the Imbalance Ratio(IR) is computed to evaluate the imbalance degree between SA and SB . It's defined as:

$$IR(SA,SB) = \frac{|sup(SA) - sup(SB)|}{sup(SA) + sup(SB) - sup(SA \cup SB)} \quad (2)$$

4.4 Rule Presentation

The rules with practical meaning are finally present as recommended routes in user interface. Apply tour route planning algorithm to rank the scenic spots or cities in a rule.

5. RESULT&ANALYSIS

5.1 Scenic spot Association Rule Analysis

Firstly, Fujian province is taken as example, in which 570 routes cover 598 scenic spots in Fujian. Here, support value is set to 0.02 and confidence value is set to 0.2. It can be found that there are six rules satisfied the criteria in Table IV. However, rule 5 and rule 6 are discarded due to their Kulc value below the threshold (Kulc measure threshold is set to 0.25). Accounting to the high IR value, rule 4 is also discarded. After applying tour route planning algorithm to the remaining rules, three recommended routes with one day, three recommended ones with two days and three recommended ones with three days in Fujian are presented on Wangluzhe website, as shown in Fig.2.More detail is seen in Fig.3, where it can be found that scenic spots in the route are arranged one by one. In the route, scenic spots have a close relationship, such as short distance, common cultural characteristics, etc.

Rule	Support	Confidence	Kulc	IR
R1:sp_1 to sp_4	0.0228	0.4333	0.4197	0.04
R2:sp_2 to sp_11	0.0210	0.2727	0.2667	0.02
R3:sp_4 to sp_3	0.0263	0.3260	0.3260	0
R4:sp_7 to sp_18	0.0403	0.5	0.4296	0.20
R5:sp_5 to sp_9	0.0263	0.3061	0.2030	0.54
R6:sp_6 to sp_12	0.0526	0.2307	0.2153	0.08

Table 4. Association rule of scenic spots in Fujian

Secondly, the rules are analyzed in Guangdong Province, in which cover 88 routes. The result is shown in Fig.4 and Fig.5. Three recommended routes with one day, three recommended ones with two days and three recommended ones with three days are included. It's clear that trans-provincial route can be recommended as well. In order to evaluate the quality of recommendation, the click number of route copy button at the top-right corner is counted. Till now, the click count is large than 100 every day.

5.2 City Association Rule Analysis

Ten provinces are taken as analysis scope, in which 876 routes cover 176 cities. The same filtering method is adopted as above. It can be discovered that several destination city has closer relationship with some departure city. As the Fig. 6 shown, when a user chooses a departure city of a route, several recommended city can be discovered in destination city list.

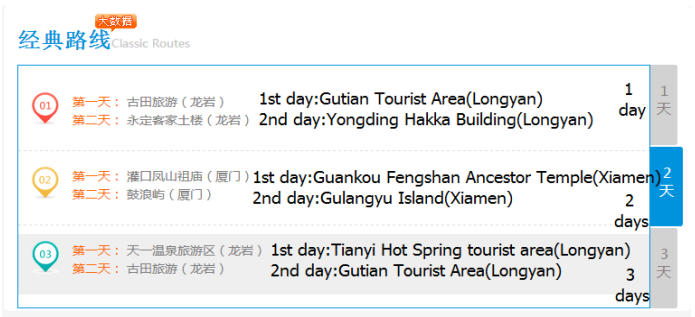

Figure 2. Route recommended routes in Fujian

Figure 3. Route detail in Fujian

Figure 4. Recommended Routes in Guangdong

Figure 5. Route detail in Guangdong

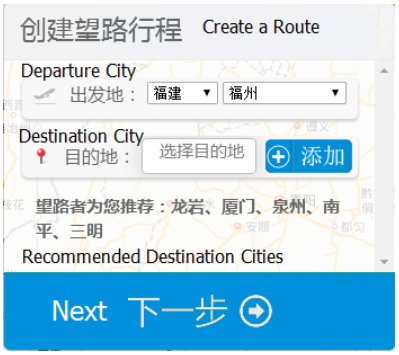

Figure 6. Recommended destination city

6. CONCLUSION & FUTURE WORK

In this paper, how to apply association rule analysis to recommend tour route to tourist is delineated and demonstrated. Here FP-growth algorithm which is combined with Kulc measure and imbalance ratio evaluation is developed and implemented. The result is analyzed and checked on WCTSNOP. It can be drawn that association rule analysis using this measure based on interestingness is effective and recommending reasonable tour route is feasible. Three different routes in Fujian are recommended for one day, two days and three days. Every route is made up of several scenic spots, whose percentage of co-occurrence is high. The route recommendations cross different cities are also found. In general, variety of routes on WCTSNOP can be recommended for diverse user needs. The quality of recommendation is evaluated by the click count every day.

However, the study still has some deficiencies. For instance, based on the model, direct search method is presented which would take up too much computing space and CPU occupation time, reducing the efficiency of rules generation. Future work will focus on the combination of tourism domain knowledge and user profile into association rule analysis for more personalized route recommendation. Further, how to incorporate other data mining method into tour personalized recommendation and improve the analysis efficiency with parallel algorithm are also in consideration.

ACKNOWLEDGEMENTS

The authors express our acknowledgement to technical staff in Fuzhou Silviscene Information and Technology Co. Ltd and anonymous reviewers for their valuable suggestions. This study is jointly supported by National Science and Technology Support Program(2013BAH28F00), Key Science and Technology Plan Projects of Fujian Province (2015H0015), and Technology Innovation Foundation of Small and Medium-sized Enterprises of Fujian Province(2015C0042).

REFERENCES

Agrawal, R., 1993. Mining association between sets of items in massive database, In: *International proceedings of the ACM-SIGMOD international conference on management of data*, pp.207–216.

Agrawal,R.,1994.Fast algorithms for mining association rules, In: *Proceedings of the international conference on very large databases*, pp.407–419.

BastideY., 2000. Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets, *Lecture Notes in Computer Science*, pp.972-986.

Chen G., 2002. Simple association rules (SAR) and the SAR-based rule discovery, *Computers & Industrial Engineering*, 43(4), pp.721-733.

Ding S., 2015. A Comprehensive Evaluation System of Association Rules Based on Multi-index, In: *International Symposium on Distributed Computing and Applications for Business Engineering and Science*, pp.304-307.

Fu X., 2000. Mining navigation history for recommendation. In: *International Conference on Intelligent User Interfaces*, pp. 106-112.

Geng L., 2006. Interestingness measures for data mining:A survey, *Acm Computing Surveys*, 38(3),pp.9-12.

Han J., 2000. Mining frequent patterns without candidate generation, *ACM SIGMOD Record*. 29(2), pp.1-12.

Han J., 2006. *Data Mining: Concepts and Techniques, Data Mining Concepts Models Methods & Algorithms Second Edition*, Wiley Online,USA,5(4),pp.1 - 18.

Hussein N., 2015.Using the interestingness measure lift to generate association rules, *Journal of Advanced Computer Science & Technology*, 4(1),pp.156.

Li Yuanbo, 2015. Recommendation based on association rules algorithm research, school of computer science, Shanxi normal university,China.

Li J., 2002. Mining Optimal Class Association Rule Set, *Knowledge-Based Systems*, 15(7),pp.399-405.

Liao S H.,2010. Mining customer knowledge for tourism new product development and customer relationship management, *Expert Systems with Applications* , 37(6),pp.4212-4223.

Lucas J P., 2013. A hybrid recommendation approach for a tourism system, *Expert Systems with Applications*, 40(9), pp.3532-3550.

Piatetsky-Shapiro G.,1991. *Discovery, Analysis, and Presentation of Strong Rules*, Knowledge Discovery in Databases, pp.229-248.

Resnick P.,1994. GroupLens: an open architecture for collaborative filtering of netnews, In:*ACM Conference on Computer Supported Cooperative Work*, pp.175-186.

Ricci F., 2010. *Recommender Systems Handbook*, Springer, pp.1-35.

Silviscene Information Technology Co. Ltd,2016. http://www.whlyw.net.

Wang D L., 2004. An Approach of Association Rules Mining with Maximal Nonblank for Recommendation, *Journal of Software*, 15(8), pp.1182-1188.

Zhou Xin., 2000. Interest degree: another threshold of association rules, *Computer research and development*, 37(5), pp.627-633.