Finding Interesting Places at Malaysia: A Data Mining Perspective

ATAE REZAEI AGHDAM, MOSTafa KAMALPOUR, DONG CHEN, ALEX TZE HIANG SIM
Faculty of Computing
Universiti Teknologi Malaysia
81300, Johor Bahru
MALAYSIA
raatae2@live.utm.my, mostafa_kamalpour_1983@yahoo.com, chendong7484@163.com,
alex@utm.my

Abstract: -This paper proposes a new opportunity for analyzing tourist profiles as found in
www.tripadvisor.com from two different aspects so as to gain conclusive results. Tourist data were “crawled”
from tripadvisor.com through our specific web crawler. A variety data mining techniques such as; clustering,
and association rules assisted to discover the first set of interesting knowledge. This was followed by
qualitative analysis applied through Nvivo software via coding of the tourist’s comments in order to propose
the design of a model. A final set of results was obtained once both results confirmed each other. In our study,
results show that there are several types of tourists with each group having different preferences. For example:
local visitors to Legoland are not satisfied with some liabilities such as the price of food. International tourists,
however, consider the affirmative points of Legoland. This research can be very useful for tourist associations
in Johor Bahru at Malaysia.

Key-Words: -knowledge discovery, data mining, tourism, qualitative analysis

1 Introduction
The World Tourism Organization estimates that
tourism is expected to grow faster than other
economic sectors in the world [1]. Uncovering new,
interesting and useful knowledge on tourist data
can be helpful for tourism organizations in order to
identify tourists’ behavior patterns and their
preferences. This paper focuses on the role of data
mining in the Malaysian tourism industry, in
particular, regarding places of interest in Johor Bahru. In essence, we will analyze tourist behavior
patterns in Johor Bahru so as to discover useful and
hidden knowledge in order to recommend
appropriate places to visit. Moreover, quantitative
and qualitative analyses have been undertaken simultaneously in this paper. Quantitative analysis
was applied by Weka machine learning software
and qualitative analysis was performed by Nvivo
software based on tourists’ comments. Each group
of analysis supports the other in order to obtain
significant results. This study can assist tourist
associations in Johor Bahru and travel agencies to
promote attractive places in Johor Bahru to both
local and international visitors.

2 Literature Review
Currently, data mining supports various kinds of
application tasks: from data pre-processing to
association rules discovery, data classification,
cluster analysis, etc [2]. Actually, it is part of the
decision-making process and the availability to
analyze data automatically helps to determine a
potential model and estimate customer behavior in
the realm of enhancing decision makers’ ability to
adjust marketing strategy and reduce the risks [3,4].
In the other study [5] Concentration was given to
to three main aspects of using data mining in the
tourism industry. These are, namely; forecasting
tourist expenditure, analyzing profiles of tourists
and forecasting the number of tourist arrivals. The
author has found various results based on these
three dimensions. For instance, in forecasting
tourist expenditure, artificial intelligence sources
such as Neural Network were used for estimating
tourist expenditure in the Balearic Islands. Also,
[6], used data mining techniques to predict
shopping expenditure by tourists with 94% accuracy. For analyzing tourist profiles [5],
categorized tourists into specific groups such as;
developmental support, prudent developers,
ambivalent, cautious and protectionist respectively
by using a clustering technique. In forecasting
tourist arrivals, some studies had been conducted
considering tourist arrivals to Hong Kong from six different countries; Artificial Neural Network (ANN) showed that this outweighed statistical methods. [5] argued that, to date, only some AI techniques such as ANN and clustering techniques have been used in tourism data mining. It is largely prepackaged software that uses these techniques readily; it can also be used for analyzing data with little training. However, the author believed that, in the future, more than one method will be applied for analyzing data. In this paper we focus on analyzing tourist profiles from two different perspectives, which includes quantitative and qualitative. In the following sections, the related studies illustrate how traditional data mining techniques have been used in the tourism industry.

2.1 Data Mining In Tourism Industry

Discovering new, stimulating and valuable knowledge by using a variety of techniques such as; classification, clustering, visualization and association rules is the main aim of data mining [7]. In the meantime, the necessity of using data mining in tourism is inevitable. In this section, we will investigate some related works regarding data mining in the tourism industry. For instance, [8], focused on using a decision tree algorithm to analyze tourist markets from two different dimensions including: impact factors of tourist spending and impact factors of tourists’ comprehensive evaluation of tourist destinations respectively. The first step focuses on total expenditure of tourists in the classification objects in the decision tree model; while the second step focuses on attributes and values including: cost, traffic, shopping environment, dining etc. By using the decision tree algorithm to analyze the data and some rules it was found that, through decision tree and associative rules mining, for example, poor shopping environments affect tourist sentiment even if it is located in an area having the best attractions. Another result illustrates that, if tourism products only focused on an attraction’s landscape in itself, the tourism product would certainly decrease the value itself. Another study proposed a new semantic association rule mining algorithm, which introduced a genetic algorithm. This method deals with textual information and divided characteristic words into various categories in an attempt to find association rules between them. They categorized tourist emergency information into five categories, namely: object, environment, activity, event and result. The author believes that this technique is useful for searching rules in a database. The results indicate, further, that the algorithm can obtain semantic rules effectively with a higher extraction rate and correctness. In extracted association rules with high confidence, the appearing words present the concepts of an emergency event. Hence, several rules discovered from small association word sets have been used to search for emergency reports of tourism. Furthermore, semantic association rules mining would obtain better result from a more sophisticated text set of tourism emergency information [9]. Data mining is present in helpful occasions of every part of tourism such as the hotel industry; for instance, by offering tourist information kiosks to provide customer information during their stay and then analyzing their data. A research study was carried out in South Korea regarding data mining; the hotel industry profiled visitors based on their preferences and comments. The author attempted to examine the effect of a customer’s demographic profile on hotel choice. Visitors were categorized by some factors including: country of origin, occupation, sex, age etc.; as a result, they identified interesting relationships existing between items in a dataset. Finally they came up with some if-then-rules, which show the probability of existence of some facts by analyzing the data. For instance, one of the rules indicates that if a customer stays at the hotel for a convention, then he/she is probably a manager. From this finding that the customer is a manager, we can acquire other information based on findings about managers. For example, if a customer prefers either a room with a single bed or a suite, pays for the hotel room by credit card, and stays at the hotel for travel/pleasure; the customer is likely to be Korean. Further, the author found that some combinations of customer characteristics have more influence on the customer’s hotel selection and patronage behavior than a single attribute. Hotel management would be wise to take into account a multitude of attributes such as: customer’s demographic profile, travel purpose, prior service experiences, as well as the availability of certain amenities [10]. It was just one of the examples among the rules that they had defined.

2.2 Techniques In Tourism Data

The most important unsupervised learning technique is that of clustering and it focuses on ways by which to identify a structure in a group or collection of unlabeled data in order to organize elements into groups [11]. Clustering refers to a
process of clustering and grouping together a collection of abstract or physical objects into similar groups. In a cluster, objects of one group are different from the objects of other groups [12]. In this paper, we intend to discover exciting knowledge, which forms the enormous amount of data in the web and by using the clustering technique. A cluster can be known as a collection of data elements, which are similar to each other and also different from other cluster elements. Clustering is grouping data into clusters with minimum similarity existing between elements of two different clusters and maximum similarity between elements of each cluster [13].

Another useful technique for discovering rules between elements is by Association Rule Mining. This technique is a data mining technique, which was introduced by Agrawal in 1993. He proposed the algorithm of Apriori in 1994 [14]. The main idea of Apriori is to find a relationship between different items of a database. The association technique is useful to find a relationship between two or more attributes and it attempts to find rules between attributes. Association rules mining helps to extract this kind of knowledge from a database so that it can be used by recommendation systems.

3 Methodology
Our methodology is categorized into five phases, which include: (i) defining required data, (ii) data collections, (iii) pre-processing, (iv) data mining and (v) conclusion. Figure 1 presents the phases of our methodology with input and output in each phase.

![Figure 1. Research methodology steps](image)

The process beginning with defining required data into tripadvisor.com then, the required data collect automatically through our specific web crawler in data collections phase. In pre-processing part, data will be ready with ARFF format to load in Wekamachine learning software and in data mining step, our dataset is analyzed from two aspects; quantitative by Weka and qualitative analysis by Nvivo. Finally, the results are summarized in conclusion step and support each other.

4 Experimental Results
Detecting significant clusters of tourist’s profiles data would be helpful for us to gain more knowledge in the realm of tourism. We executed the K-Means and EM algorithms in order to determine the existence of some clusters from visitor data. After clustering our dataset, we discovered that some clusters were important for us by which to gain more knowledge. Table 1 illustrates a summary list of findings from the clustering section.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Malaysian visitors aged between 35 to 49</td>
</tr>
<tr>
<td>2</td>
<td>Singaporean tourists aged 18 to 24.</td>
</tr>
<tr>
<td>7</td>
<td>Great group of Singaporean female visitors.</td>
</tr>
</tbody>
</table>

4.1 Interesting Rules Amongst Tourist’s Data
By escalating the number of attribute findings, identifying the rules between them becomes an increasingly difficult task. However, some rules are not intuitive in a tourist’s dataset. In this instance, association rules mining techniques can lead to the appearance of some hidden rules among items in the dataset. An Apriori technique is then applied to uncover hidden relations between elements in the dataset. Amongst numerous rules, we were able to discover some useful and valuable ones, some of which are listed below:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If a tourist is from Australia, female, and aged 35 to 49, then she will most likely select Legoland as a fun travel destination.</td>
</tr>
<tr>
<td>2</td>
<td>If tourists are aged 25 to 34 and choose Sanrio Hello Kitty Town, they are most probably females.</td>
</tr>
</tbody>
</table>

These kinds of findings were not detected through clustering techniques; therefore, association rules lead us to gain more interesting knowledge from tourists’ respective datasets. In the next step, we proceed to analyze the same dataset using qualitative analysis by Nvivo. Through the
application of Nvivo, the comments of tourists will be analyzed to catch the salient points of comments made. We hope to support both findings using different approaches.

4.2 Textual Analysis of Tourist’s Profile

The Nvivo qualitative analysis was carried out to analyze the comments tourists tend to make about various places. Using Nvivo actually helped us to make our qualitative research more transparent; it provided us with the ability to better address some of the key challenges in professional communication research, including: efficiency, multiplicity and transparency [15,16]. We coded tourists’ comments to extract a model based on some attributes. The figure below is the output model of Nvivo achieved after coding tourists’ comments based on the impact factors of Legoland and the country of origin of visitors.

As can be seen in Figure 2, the impact of Legoland as extracted from visitors’ comments can be disseminated into two points, namely: positive and negative. Affirmative points argue the need for a suitable place to be enjoyed by the whole family, including children; while the negative points present liabilities of Legoland, such as food prices and hot weather. A number of results can be extracted from this model. For example, Australian tourists concentrate on positive points such as it’s being a fun place for kids and great for the family; whereas Indonesians argue that the price for food is expensive. What is interesting in this figure is, in spite of there being negative points; some foreign visitors (notably Singaporeans and Australians) only considered the positive points. Likewise, word frequency can be achieved through Nvivo for identifying the most frequently used words and their synonyms in tourists’ comments. In this case, the results indicate that a great bulk of visitors emphasize positive points of Legoland; which was the same result as in the model. Nvivo analysis supports our result in the Weka analysis section; therefore, these two analysis techniques can be useful in the tourism industry in order to acquire definite results of tourist data.

5 Recommendation for Tourists

According to the results provided in previous sections, we can recommend a list of appropriate places and services for specific groups of tourists based on their preferences. For places of interest, we can suggest the model below for visitors. This model demonstrates suitable places for each group of tourists; for instance, Legoland is one of the best options for Malaysian, Singaporean and Australian travelers.

Figure 3. A recommendation model for interesting places to visit

As can be seen in Figure 3, the recommendation model for interesting places to visit includes Legoland and Danga Bay. According to the findings in the association rules section in Table 2 and the Nvivo model (Figure 2), we conclude that Legoland can be included on our recommendation list for Australian female tourists, as well as Malaysian and Singaporean adults.

6 Conclusion

This study has applied quantitative and qualitative analysis together in an attempt to discover hidden knowledge from tourist profiles in www.tripadvisor.com. Several data mining and textual analysis techniques (such as clustering, association rules mining and Nvivo qualitative analysis) have been carried out to show the performance of data mining in the tourism industry. In essence, we have shown that Nvivo analysis in part can support the findings of the Weka analysis section of our research to gain conclusive results. Findings of this study can be used by touristic association in Johor Bahru in order to identify tourist behavior patterns, improve their services, offer specific travel packages and, finally, enhance the number of tourists arriving in Malaysia.

Acknowledgement

The author, AtaeRezaeiAghdam, would like to express gratitude for the thoughtful suggestions and warm support received from the postgraduate students in the field of IT Management in...
UniversitiTeknologi Malaysia (UTM). He would also like to thank his dear parents for their love and encouragement. UTM is gratefully acknowledged for its Research University Grant (vot 07J79).

References: