Uncovering Hotel Guests Preferences through Data Mining Techniques

Mostafa Kamalpour[1†] , Atae Rezaei Aghdam[2††], Shuxiang Xu[3], Ehsan Ghasem khani[4] Aryan Baghi[5]

[1]Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia,
[2 & 4]Department of Computer Science, University of Applied Science and Technology, Center of Tehran municipality ICT org, Tehran, Iran,
[3]Faculty of Science, Engineering and Technology, School of Engineering and ICT, University of Tasmania, Tasmania, Australia
[5]Faculty of Computer Engineering, University of Science and Technology (IUST), Tehran, Iran

Summary
The proliferation of online travel communities, travel websites, and technology developments are driving tourism industry to develop new methods for marketing and improving customer satisfaction. The main aim of this study is to analyze the potential use of Data Mining and Web Mining techniques in tourism industry to extract the hidden knowledge from hotel visitors' information. For this purpose we have collected the data, from visitors of Mersing Island hotels as found at www.tripadvisor.com through our task specific "RK" web crawler, which collected 616 user profiles information. The research method used in this research is CRISP DM, and by using this method along with "RK" Crawler, two models have been proposed to use by managers in order to improve customer satisfaction. Results show that there are a various type of tourists with each group having different preferences. For instance, if a visitor is from Singapore, male, and interested in great foods and wine, he is also interested in outdoor and adventure activities. This research study can be very helpful for tourist association, hospitality, and hotel managers.

Key words:
Web Mining, Tourism, Data Mining, Hotel.

1. Introduction

Over the last decade, an archetype change has occurred in business computing. It came with the emphasis of moving from approach of collection of data to knowledge discovery. During this change explosive growth of the World Wide Web has been happening, which has made many technologies. These advances have enhanced data collection frameworks and eventually resulted in new techniques for knowledge discovery from huge databases. One of the successful and popular techniques for extracting knowledge from web data is web mining [1]. In fact web mining is a kind of data mining for web data. For example, it enables businesses to turn their huge amount of transactional and Website usage data into the actionable knowledge, which is useful for every level of the organization, and not only for the front-end of online stores. Data collection, data exchanging and exchange of information became easier by rapid growth of the World Wide Web, enabling technologies, and having resulted in speeding up of highest major functions of businesses. Today some problems such as delay in manufacturing, shipping, retail, or customer service processes are not counted as necessary evils anymore, and companies enhancing upon these essential functions have an edge in their battle of margins. Technology has been brought to put up with countless business processes and affected huge change in the form of communications, tracking, and automation, but many of the most profound and unpredictable changes are yet to come. Leaping in computational power has enabled businesses to collect, and process large amounts of data. The accessibility of data and the essential computational resources, with the potential of data mining all together, has shown great promise in having an excessive transformational effect on the way businesses perform their work. Some companies used advantages of this end such as Amazon.com as evidence, leveraging of large repositories of data collected by firms will cause data mining techniques and methods offer unparalleled opportunities in understanding business processes and in prediction of future behavior. The proliferation of online travel communities and sharing travel experience amongst travellers and the profiling of tourists has become a major issue [2]. In this paper we investigate the role of Web mining in tourism industry specifically hotels in Mersing island in Malaysia. In fact, we present that how web mining can increase the efficiency of hotel's performance, what the preferences of the visitors from different countries are, and what the hidden and profound hints are, which can enhance tourism industry's profit and customer satisfaction as well. This paper is organized in the following manner; section 1; study of the related work in this discipline. Section 2; description of our research methodology and pre-processing steps and lastly, section 3 illustrates the

Manuscript received August 5, 2017
Manuscript revised August 20, 2017

experimental results along with recommendations for tourism industry.

2. Literature Review

Rapid growth of Internet users indicates that users accept online transactions widely. If a hospitality firm were able to provide functional and valuable web sites, which attract their visitors, definitely they will also get benefits ultimately [3]. Actually, most effective websites are those that give customers the easiest access to relevant and useful information. Two thirds of travelers have the experience of using online reservation. A research on Malaysian hotels shows that among 3, 4, and 5stars hotels, 5star hotels are using the benefits of the Internet more than the rest. The variables that used to measure the results are general presentation of the website, information and service, technology and interactivity, and finally promotion and marketing [4]. Another section of the research revealed that the website performances were not what travelers expected. Furthermore, Assessment of 228 Spanish companies discovered a positive relationship between external web contents and company's performance [5]. In addition, Internet is using as sales and marketing tool in the hotel and tourism industry. A model had designed to analyze the variables and measure them in order to elicit valuable knowledge. The study was based on 52 hotels in Taiwan and they measured the mentioned factors, and analyzed them in which to be understandable by representing in statistical form [6]. Using web data extraction to develop innovative web information systems by relying other websites information can be as a thrust. ImBored.com is a Malaysian website which collects data from different sources in the web which are related to the hotels and tourism. It presented itself as a very useful portal to tourists. By using Internet or GPRS, tourists are able to collect information in one site only. They are also able to collect data about accommodation, channel listing and movie updates. An important issue is that extracting and retrieving data from several different web sources through web wrapper engine each time is time consuming. Concurrent use of the engine is still to be tested for its efficiency. Effective and efficient use of databases as a caching mechanism is complementary approach, which can be considered in the future. In addition, when we deal with a high number of heterogeneous web sources, we need to address the issue of semantic conflicts with respect to data interpretation [7]. Another research has done in South Korea regarding data mining and hotel industry. The authors tried to profile visitors by the data that they normally leave after their residence. They categorized visitors by many factors such as country of origin, occupation, sex, age, and so on. Hence, they could find very nice relationship between the raw data that they had and finally they came out with some if-then-rules, which show probability of existence of some facts by analyzing the data. For instance one of the rules indicated that if a customer stays at the hotel for a convention, then he/she is probably a manager, and from this finding that the customer is manager we can get much other information based on findings about managers [8]. Another research has done in South Korea to investigate visitor preferences. A survey distributed among 281 customers helped to get closer to the answers of some questions, such as which customers are likely to return to the same hotel? Or which service attribute is more important to customers and so on. The survey took place among 281 guests and 11 hotels in Seoul, South Korea. Many characteristics collected and considered as sources of data such as price range, location, and service amenities. Visitors also provided the authors with data which was related to their demographic profile such as age, occupation, gender, nationality and so on. All participants had visited at least one of the eleven hotels [9]. Furthermore, one research study attempted to find the level of satisfaction of travelers with Malaysian hotels. Their research question is to find differences between Asian and Western travelers in evaluation of customer satisfaction with Malaysian hotels. The goal of the study was to suggest some ways to improve the hotel services, and outline factors. By distributing questionnaire among 200 tourists in KLIA airport, the data has collected and analyzed. Result claimed that 43% of visitors were in age group of 21-35 and 32% of them were in age group of 36-50. There were some attributes in questionnaire such as cleanliness, rooms, location, sleep quality, value for money, and services [10]. Similar research has done by Choi and Chu concerning the Asian and Western travelers in the Hong Kong. The method of data collection was questionnaires at the Hong Kong airport, spreading among those visitors who leave the country. Within the questionnaire there were some factors that asked from travelers such as demographic and personal questions, and also some questions regarding hotels that visitors used to reside. The factors relating to hotels included cleanliness, location, room rate, service, value for money, and location. 540 people answered to the questionnaire and the collected data was analyzed using factor analysis. Result illustrated that Asian travelers spend less money than Westerns. One of the reasons for this issue was that most Asian countries were among developing countries that normally they have less salary compare with Westerns. Analysis of data also showed that Westerns spend 45% of their budget on accommodation whereas Asians spend less than 25% for the accommodation. Another result, which was supporting previous one, was that 70% of Asians were looking for midrange hotels. Final result claimed that Asians were interested in shopping more than Westerns. Asians likes to spend more than 50% of their budget on shopping. Asians were also caring more about value for money, but

Westerns were looking for room quality. The final result specified that Asians were very interested in spending most of their time on entertainment during the trip [11].

Furthermore, Richard S.Segall and Qingyu Zhang applied techniques of web mining to actual text of comments that were written by hotel visitors by using Megaputer PolyAnalyst. Megaputer PolyAnalyst is an enterprise analytical system, which has unique feature of integrating data and text mining together with web mining. The software included functions such as clustering, keyword and phrase extraction, link analysis, dimension matrices, and taxonomy. As a result, they found many rules based on their data collection and analyzing of them [12].

Another study reviewed the foundations of satisfied and unsatisfied hotel customers. A text-mining method was followed and online reviews by satisfied and dissatisfied customers were compared. Online reviews of 2,510 hotel guests were gathered from TripAdvisor.com for Sarasota, Florida. The research outcomes discovered some common classes that are used in both positive and negative reviews, including place of business (e.g., hotel, restaurant, and club), room, furnishing, members, and sports.

Study results indicated that satisfied customers who are willing to recommend a hotel to others refer to intangible aspects of their hotel stay, such as staff members, more often than unsatisfied customers. On the other hand, dissatisfied customers mention more frequently the tangible aspects of the hotel stay, such as furnishing and finances [16]. Likewise, In a study regarding exploring the influence of online review on traveler's preferences, researchers proposed a model by harvesting data from business travelers in Mainland China. Based on experimental results, six features of online reviews content and one source attribute were recognized, namely, usefulness, reviewer expertise, timeliness, volume, valence (negative and positive) and comprehensiveness. Results also affirmed positive causal relationships between usefulness, reviewer expertise, timeliness, volume and comprehensiveness and respondents' online booking intentions. Additionally, considerably negative relation between negative online reviews and online booking intentions was explored, however impacts from positive online reviews upon booking intentions were not statistically significant [17]. Another research investigated the effectiveness of big data analytics to better understand vital hospitality issues, specifically the connection between hotel visitors experience and satisfaction. Specifically, this study carried out a text analytical approach to a large quantity of consumer reviews mined from Expedia.com for analyzing hotel guest experience and inspect its relationship with satisfaction ratings. The findings exposed various proportions of guest experience that carried varying weights and, more outstandingly, have novel, meaningful semantic compositions. This study showed that big data analytics would produce new insights into variables that have been comprehensively studied in existing hospitality literature [18].

A comparative analysis is done by examining primary and secondary sources to explore which source of information provides the most attractive data to be mined in tourism field by analyzing the searching ability of the tourist destination through a set of pre-defined keywords. Further to a collection of the sources and their analysis, authors have compared them by considering the volume of tourist searched relevant keywords. The study illustrated that the highest impact on the searching ability of a tourist destination has on the tourist location is virtual communities and reviews found on the Internet [19].

Another research study founded the progress of information and communication technology (ICT) based on a review of papers published in tourism and hospitality journals between 2009 and 2013. 107 journal papers were retrieved and evaluated. The papers were categorized into two major groups, consumer and supplier, which mostly comprise the key players in the industries. A content analysis revealed that hospitality and tourism industries use ICT in different functional units and for different applications [20]. According to sentiment analysis, one study has analyzed 3,000 tweets by local residents and 3,000 tweets by tourists at 10 main destinations in Europe with to find out their divergence. The consequences presented that, the tweets by both local residents and tourists tend to be positive. Nevertheless, some destinations have high percentages of neutral and even negative tweets [21]. Furthermore, a research surveyed 150 female working adults in Malaysia, to conclude that single business female travelers tend to be younger, unmarried and engage in personal leisure activities. According to experimental results, security and location are the main concern issues while they are selecting a hotel. They concluded that female business traveler requests must be accommodated if hotels establishment with to attract this growing market [22]. Researchers in another study mined characteristics from two samples of five-star hotel reviews in Jakarta and Singapore through text mining methodology. The other purpose of their study was to explain locating of five-star hotels in Jakarta and Singapore based on the obtained attributes using Correspondence Analysis. They contended that reviewers of five star hotels in both cities mentioned similar attributes such as service, staff, club, location, pool and food. Attributes derived from text mining seem to be viable input to build fairly accurate placing of hotels [23].

Another research offered a text-mining technique to apply in marketing strategies for discovering critical features in the hospitality sector from customers' online reviews by extending Herzberg's Two-Factor Theory of Motivation. The results provided the list of words including how many times they occurred in satisfaction and dissatisfaction reviews. Significant ones are selected, and after

conducting a correlation analysis, they were categorized into three classes: (1) motivators: words correlated with total review scores only in the positive reviews; (2) hygiene factors: words negatively correlated with total review scores only in the negative reviews; and (3) effective motivators: words correlated with total review scores in both positive and negative reviews [24].

In a research done by Barnes and Jia, they explored the significant aspect of customer service voiced by hotel quests use a data mining method, latent dirichlet analysis (LDA). The gathered 266,544 online reviews for 25,670 hotels located in 16 countries. LDA discovered controllable magnitudes that were the main for hotels to manage their interactions with tourists. Result of the study disclosed that budget hotels usually offer a partial hotel service and their rates are 25-30% cheaper than average market rates. Most hotels in this segment are considered zero to three-star, but, apart from price, there was a shortage of understanding about the factors affecting customer behavior. Based on an online review analysis, several important non-price scopes are recognized by visitors of two to three-star hotels, comprising “bathroom” and “checking in and out”. Research outcomes advised that 5-star hotel managers must focus upon the feeling homeliness for quests; it is one of the most significant measurements motivating customer satisfaction for 5-star hotels, especially compared to other lower-level hotels. LDA analysis presented that price is important factor for men than women then, hotel managers might take offering some special discounts for single male customers into account, specifically during low peak periods. Likewise, older hotel customers value homeliness more than younger customers [25]. With aim of building long-term oriented profitable relationships with customers a study is done by Dirsehan to clarify the use of text mining through Rapidminer software. As a result, an overall guideline provided for researchers and practitioners, to enable them to perform basic text mining by themselves [26].

3. Methodology

In this research study Cross Industry Standard Process for Data Mining or CRISP DM, which is one of the most common methodologies has been applied. Cross Industry Standard Process Model for Data Mining (CRISP-DM) was developed as an open standard by leading KDD appliers and a tool supplier [13]. This model was established in the 1990s, and is a data mining process model for data mining experts [27]. CRISP DM is developed by a consortium led by SPSS, and SEMMA, developed by SAS [29]. The current CRISP-DM Process Model for KDD provides an overview of the life cycle of a KDD project in figure 1. It contains the corresponding phases of a project, their respective tasks, and relationships between these tasks.

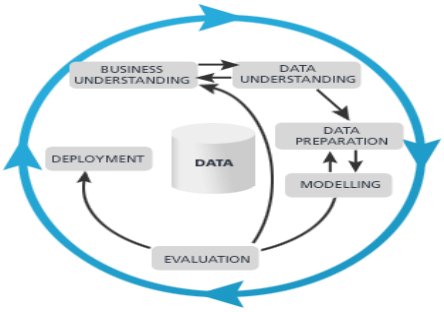


Figure 1. Data mining methodology

As it is shown in the figure 1, this methodology contains some steps including business understanding, data understanding, data preparation, modeling, evaluation, and deployment [14][30]. The main aim of CRISP DM is to provide common mechanism, procedures and recommendations for each data mining steps. The steps of CRISP model for web mining in Mersing Island are explained as below;

3.1 Business Understanding

Business understanding is focusing on project objectives and requirements from business perspectives that are further converted into data mining goals or problem definitions. Plan for the whole process is created based on identified inputs and specified goals.

3-2 Data Understanding

The data understanding covers basic operations with initial data collections mainly oriented to get familiar with them, to evaluate and assess the data quality or to obtain basic characteristics and statistics of the investigated historical dataset. This phase is started with the selection of data relevant for the specified problem. We have investigated the following data sources: customer’s profiles inside the website (www.tripadvisor.com), for those who had visited Mersing Island, all Mersing’s hotel’s information based on saved data in website, and reviews of customers who had visited these hotels from same website.

3-3 Data Preparation

The data preparation represents all activities to construct the final dataset for modeling purposes. This is a very important phase for the whole process and expected results and in many cases has several iterations. Relevant tasks include table, record, and attribute selection as well as transformation and cleaning of data based on selected modeling algorithms and their conditions. This phase is consists of “Data Extraction” from the website which is storing in MySql database and as Excel files as well. We collected the data and saved them into our database by using RK Crawler, which is software that specifically developed to extract data from www.tripadvisor.com.

Using this program is simple. In this research study 616 tourists profiles are collected through this specific web crawler.

3-4. Data Modeling

In the data modeling phase, various algorithms and techniques are selected and applied on prepared data. Important step represents calibration of algorithm's parameters to optimal values based on obtained results. In this phase several main data mining techniques can be used such as prediction, classification, clustering, association rules, and etc. Classification technique is not using because it is applied for predefined classes and in this project we are not aware about these classes and we intend to find clusters that their instances have some similarities which is not possible by classification. An association rule also used to find the knowledge from our data. Association rules has several advantages that make it the most accurate technique for mining datasets and finding rules within them. Implementing the association rules is easy and this technique has a great ability for hiding sensitive data [15]. In this research srudy, we use Apriori algorithm. Apriori is a powerful technique that inevitably depends on two main parameters of minimum support, and minimum confidence.

3-5. Evaluation

In this phase, several evaluation steps are performed: all created models are interpreted and evaluated with respect to the specified business objectives; the creation processes reviewed step by step. At the end of this stage, a decision on the use of data mining results should be reached. The first method is use of Microsoft Excel and the second approach is analyzing the data using data mining techniques. Clustering and association rules have been carried out as two techniques of data mining.

3-6. Deployment

The deployment phase contains presentation or report describing obtained results for customers with proposed deployment steps. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. But the collaboration between both of them is important for effective deployment and successful realization of the whole data mining process.

3-7. Data Preparation for WEKA software

In this study we carry out Weka machine learning software to extract the hidden knowledge from data. WEKA is a data mining toolkit that offers a standard format for running machine learning algorithms [28]. Weka use ARFF (attribute- relation file format) as its input file and in many cases preparing the input file needs much more effort. We used different types for attributes such as numeric, nominal, and string. Numeric is using for numbers, nominal for predefined values, and string for strings. These types are shown like below:

@RELATION <Comment_Table>
@ATTRIBUTE Visitor_Id numeric
@ATTRIBUTE Visitor_Profile_Name
@DATA
1,Mostafa
2,Ahmad
3,John
4,Mohammad
5,Joe

Comment_Table is the name of database and the code above is calling the database to be read in Weka. The next tag is meaning that there is an attribute within the dataset with name of Visitor_Id and it is number. The last tag or @DATA, is start point of data that we have in database. The important point to avoid errors is to have compatible data in dataset based on the attribute type that we defined earlier.

4. Analysis and Findings

The process of analyzing tourist data in this paper started with gathering data from www.tripadvisor.com through our "RK" crawler, which is task specific web crawler. This crawler is able to accumulate the hotel name and search the tripadvisor database for this name.

4.1 Statistical perspectives of tourist data

In this section, Microsoft Excel analyzes the tourist data from statistical perspective. The analysis is done based on the reviews about 18 hotels, and guest houses in Mersing Island. Table 1 shows the list of hotels and their codes that we assigned to them.

Table 1. Hotels By Their Codes

Hotel ID	Hotel Name
1	Sea Gypsy Village Resort & Dive Base
2	Rawa Island Safaris Resort
3	Embassy Hotel
4	Alang's Rawa
5	Hotel Havanita
6	Hotel Timotel
7	Sibu Island Resort
8	Fishing Bay Resort
9	Raja Villa Resort & Hotel
10	Seri Malaysia Mersing
11	Omar's Backpacker's Hostel
12	Teluk Iskandar Inn
13	Batu Batu
14	Felda Residence Tanjung Leman
15	Kali's Guesthouse
16	Riverside Hotel
17	Sari Pacifica Hotel, Resort & Spa - Sibu Island
18	D'View Hotel

As it can be seen, Table 1 presents the frequency of reviews regarding each hotel sorted by hotel codes. Hotel

with code two has the most number of reviews with 122 reviews and of course it has the most reliable data, and hotel with code 18 has the least with only one review.
The process of finding average review rate is based on the six factors of value, rooms, service, sleeping quality, location, and cleanliness. Each of these six factors is rated between one and five. Result of this comparison is showing that the average of review rate is completely depended on value. If the value is high, probably the review rate is also high, if the value is low; review rate will be low as well. Figure 2 compares all factors with each other with association of statistic about all hotels based on the factors. Factor of sleep quality also seems independent from other four factors because in most instances it is ranked higher than other factors.

In fact, there are three factors that their average value is less than the average review of hotels. Factor of cleanliness has average of 3.42 where the factor of value is 3.40 and service factor is 3.37. Hence based on the statistics, it is the most problematic area which voted by visitors are cleanliness, and services which must be improved.

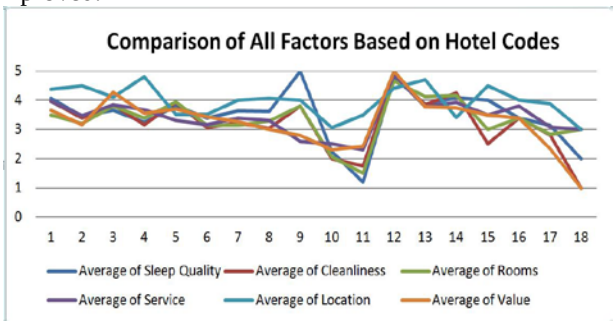


Figure 2 . Comparison of All Factors Based on Hotel CodesWith Full Statistics

Among all the countries, we have chosen 16 of them that have five or more visitors to increase the preciousness the result. Figure 3 compare these visitors based on their country of origin and according the average review rate, and three problematic factors of cleanliness, service, and value. Results indicate that Australian is happiest group among all visitors and the average review rate of them is 3.79 and respectively British with 3.77, and Indonesian with 3.71. But unhappiest visitors are those who are from Hong Kong with average review rate of 2 and after them are Italian with rate of 2.25, and Chinese with 2.77. The reason for these differences is definitely variety of expectation and customization based on their standards and life styles.
In terms of cleanliness, tourists of Hong Kong, Italy, and Switzerland are unhappiest, and visitors from United Kingdom, Indonesia, and Australia are most satisfied. The column of the services reveals that visitors from Hong Kong, China, Italy are not satisfied and they ranked as unhappiest in terms of hotel services. However, British, Indonesian, and Australian are satisfied with the service level of the hotels.

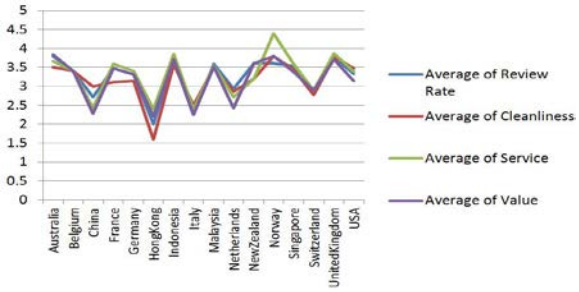
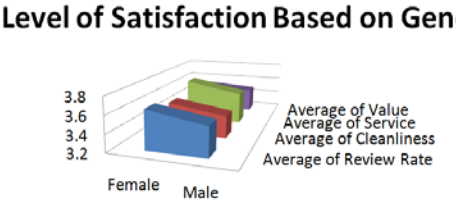


Figure 3 . Review Rates Cleanliness, Services, and Value Based on Country of Origin

Expectations of females and males are normally different. Figure 4 illustrates the differences of male and female visitors based on the rates of the services, cleanliness, reviews, and value. The result shows that in all three factors, male are less satisfied compared to females. Therefore, managers must have better services to male visitors due to their higher expectations.



	Female	Male
Average of Review Rate	3.648648649	3.538461538
Average of Cleanliness	3.556756757	3.455621302
Average of Service	3.691891892	3.568047337
Average of Value	3.518918919	3.479289941

Figure 4. Level of Satisfaction Based On Gender

Based on the figure, the result for age group of 25 to 34 is totally different. Males in this group age are more satisfied than females according to the average review rate and also based on the three factors. For the factors of cleanliness, service, and value males are more satisfied than females, but the gap is small. One of the big groups of tourists is age group of 50 to 64. The uniqueness of this group is that same as age group of 18 to 24, males are less satisfied with every factor than females. Figure 5 depicts all the factors for this particular age group of 65+.

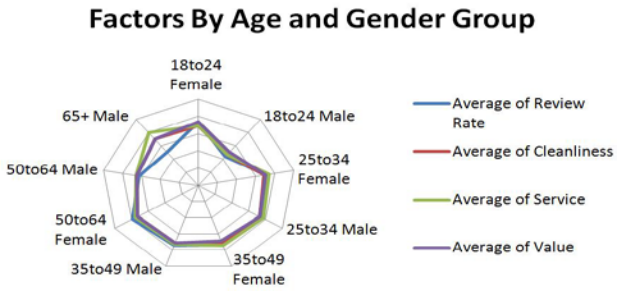


Figure 5. Factors by Age and Gender Group

4.2 Interesting rules relating to tourist data

Two techniques of association rules and clustering are used in this research in order to discover valuable knowledge. In case of visitor interests, Different people have different preferences and as we found in visitors profiles, each of them wish to have some kinds of fun in their trip. Table 2 is showing all interests of visitors.

Table 2. Visitor Interests

Number	Interest
1	To visit beach
2	To visit museums or historical places
3	To visit theme or amusement parks
4	Outdoor or adventure
5	To visit casino and doing gambling
6	To have great food or wine
7	To go shopping
8	To visit spa centers
9	To play golf
10	To do winter sports
11	To visit music festivals
12	To attend sporting events

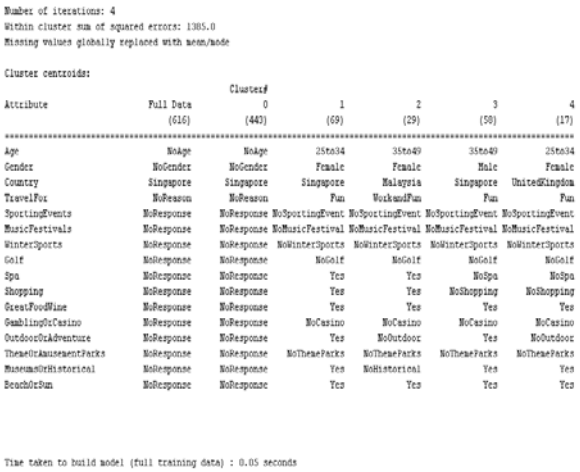
Every person has his/her own interesting activity to include in his/her trip. According to our data, visiting beaches is the most favorite activity of visitors. 162 visitors mentioned their interest of visiting beaches within their profile. There are two activities of playing golf, and gambling, which are not favorable. The frequency of interest attributes present in below table;

Table 3. Frequency of interests

No	Interest	Yes	No	Total	Interested Respondents
1	To visit beach	162	11	616	93.64 %
2	To have great foods or wine	138	35	616	79.76%
3	Outdoor or adventure	124	49	616	71.67 %
4	To visit museums or historical places	122	51	616	70.52 %
5	To go shopping	85	88	616	49.13 %
6	To visit spa centers	70	103	616	40.46 %
7	To visit music festivals	40	133	616	23.12 %
8	To visit theme or amusement parks	38	135	616	21.96 %
9	To do winter sports	34	139	616	19.65 %
10	To attend sporting events	26	147	616	15.02 %
11	To visit casino and doing gambling	5	168	616	2.89 %
12	To play golf	4	169	616	2.31 %

B1. Clustering by using simple K-Means

By using Simple K-Means algorithm, we cluster our data into five clusters. As a result presenting in Figure 6, first cluster is not having any valuable information due to the large amount of NoResponce data in dataset, but the other four clusters are exactly showing the findings that we had in previous steps.



Number of iterations: 4
Within cluster sum of squared errors: 1385.0
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (616)	Cluster# 0 (443)	1 (69)	2 (29)	3 (58)	4 (17)
Age	NoAge	NoAge	25to34	35to49	35to49	25to34
Gender	NoGender	NoGender	Female	Female	Male	Female
Country	Singapore	Singapore	Singapore	Malaysia	Singapore	UnitedKingdom
TravelFor	NoReason	NoReason	Fun	WorkandFun	Fun	Fun
SportingEvents	NoResponse	NoResponse	NoSportingEvent	NoSportingEvent	NoSportingEvent	NoSportingEvent
MusicFestivals	NoResponse	NoResponse	NoMusicFestival	NoMusicFestival	NoMusicFestival	NoMusicFestival
WinterSports	NoResponse	NoResponse	NoWinterSports	NoWinterSports	NoWinterSports	NoWinterSports
Golf	NoResponse	NoResponse	NoGolf	NoGolf	NoGolf	NoGolf
Spa	NoResponse	NoResponse	Yes	Yes	NoSpa	NoSpa
Shopping	NoResponse	NoResponse	Yes	Yes	NoShopping	NoShopping
GreatFoodWine	NoResponse	NoResponse	Yes	Yes	Yes	Yes
GamblingOrCasino	NoResponse	NoResponse	NoCasino	NoCasino	NoCasino	NoCasino
OutdoorOrAdventure	NoResponse	NoResponse	Yes	NoOutdoor	Yes	NoOutdoor
ThemeOrAmusementParks	NoResponse	NoResponse	NoThemeParks	NoThemeParks	NoThemeParks	NoThemeParks
MuseumsOrHistorical	NoResponse	NoResponse	Yes	NoHistorical	Yes	Yes
BeachOrSun	NoResponse	NoResponse	Yes	Yes	Yes	Yes

Time taken to build model (full training data) : 0.05 seconds

Figure 6. Five Clusters by Using Simple K-Means Algorithm

B.2 Applying association rules by using Apriority Algorithm

After applying Association rules techniques on our dataset the results are summarized in below table;

Table 4. Results of Association Rules Technique

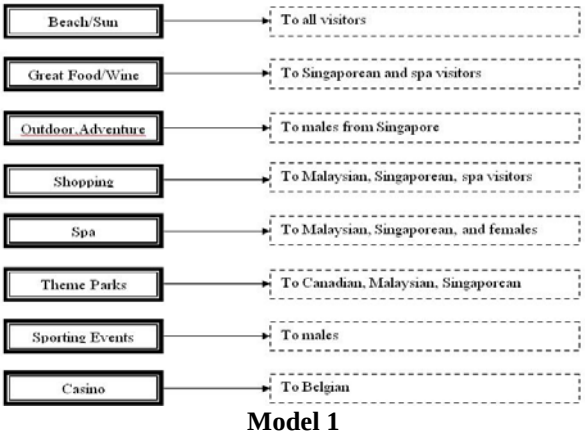
Result
79 % of visitors are interested to have great foods and wine during their trip.
71 % of visitors are interested to have outdoor and adventure activities during their trip.
70 % of visitors are interested to visit museums, historical, and cultural places during their trip.
22 % of visitors are interested to visit theme and amusement parks during their trip.
20 % of visitors are interested to experience winter sports during their trip.
15 % of visitors are interested to attend sporting events during their trip.
3 % of visitors are interested to do gambling and visit casinos during their trip.
Biggest group of visitors who are interested in shopping are Malaysian and Singaporean.
Only Belgian, Singaporean, and Malaysian visitors are interested in gambling.
Visitors of Germany, France, Indonesia, and Netherlands are mostly male.
Visitors from Canada, Belgium, Finland, Malaysia, and Singapore are more interested to visit theme and amusement parks.
Males are more interested to visit theme and amusement parks compare with females.
Visitors in clusters 1 and 3 are interested in outdoor and adventure activities.
Visitors in clusters 1, 3, and 4 are interested to visit museums, historical, and cultural places.
Females from Singapore are more interested in spa and shopping than males from Singapore.

Males in age groups of 18-24 and 50-64 are not happy with Mersing hotel's services.
Males in all age groups are less satisfied than females about Mersing hotels.
Visitors from Italy, Hong Kong, and China are less satisfied with Mersing hotels compare with visitors of other countries.
Location factor is more satisfying than other factors for all types of visitors.
Most Mersing visitors are from Singapore with 214 visitors, Malaysia with 99 visitors and UK with 61 visitors.
Singaporean visitors, who like to go to spa, and shopping, are interested in great foods and wine as well.
If a male in age group of 25-34 visit beach, he likes outdoor and adventure activities as well.
Visitors from UK, who go to beach, are interested in great foods and wine as well.
If a visitor of spa is in age group of 25-34, he/she is interested in great foods and wine as well.
Singaporean spa visitors are interested in great foods and wine.
If a visitor is from Singapore, male, and interested in great foods and wine, he is also interested in outdoor and adventure activities.

Mersing Island is potentially place for attracting visitors due to the natural beauties. Our result shows that most visitors who visit this island are from Singapore, Malaysia, and UK. There are high interests in different activities and the result shows that visiting beaches and historical places, having great foods and wine, and also having outdoor and adventure activities are very interesting for visitors. Visitors from different countries also have different preferences. For example, those who are from Malaysia and Singapore are very much interested in shopping and spa.

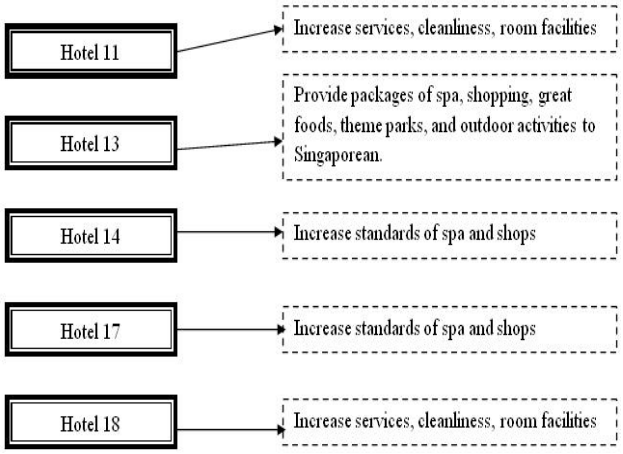
4.3 Recommending a policy for hotel managers

According to the results based on analyzing data through data mining techniques by WEKA software, we list a policy for hotel managers to offer special package for specific group of tourists based on their activities.



Model 1

This model is proposed to use by hotel managers. It is contents of 8 rules that we have found based on our analysis. For instance, the first rule asks managers to offer beach/sun packages to all travelers because the finding revealed that 93% of travelers are interested in this kind of activities. Below model is showing a model, which is recommending some particular hotels to increase their service based on some shortcoming, which have found in our results.



Model 2

By summarizing the rules, which extracted from association rules mining, we suggest five rules regarding hotels. The first one indicates that hotel with ID number 11 needs to increase level of services, cleanliness, and room facilities to satisfy its customers. Singaporeans are very interested in spa, shopping, great foods, theme parks, and also outdoors activities, and because many Singaporeans are residing in hotel with ID 13, so they need to increase their services from these aspects.

5. Conclusion and Future Works

Choosing a tourist destination to travel is one of the hectic activities due to numerous choices travellers usually come across. There are various sorts of factors which should be taken into account, ranging from activities to transportations and so on. In this study, we have applied data mining techniques such as visualization, clustering, and association rules to extract the hidden knowledge from tourist profiles on www.tripadviros.com in the area of Mersing Island. In essence, we have shown that some factors play an indispensable role in tourist hotel preferences. Findings of this study can be used by tourist associations and hotel and hospitality managers in Mersing island in order to develop their hotel and room services, offering a specific packages for special group of travellers and finally, increase the number of visitors arriving in Malaysia.

Future works can be focused on large scope for the mining of voluminous amount of data, extracting valuable knowledge for all tourists' spots in Malaysia. Accordingly, it would be beneficial to propose an automated process by which to collect and prepare data with suitable format autonomously. In this respect, the authors would not have to pre-process and convert the format manually. Crawled data would actually be ready to be analyzed through just a few clicks in a short space of time. Hence, it would provide an opportunity to analyze the tourist profile of all interesting spots and hotels in the world in a short time-frame and easy manner.

References

[1] Rao, R. S., & Arora, J. (2017). A Survey on Methods used in Web Usage Mining.

[2] Aghdam, A. R., Kamalpour, M., Chen, D., Sim, A. T. H., & Hee, J. M. (2014, August). Identifying places of interest for tourists using knowledge discovery techniques. In Industrial Automation, Information and Communications Technology (IAICT), 2014 International Conference on (pp. 130-134). IEEE.

[3] Wu, F. (2010, March). Notice of Retraction Apply Data Mining to Students' Choosing Teachers Under Complete Credit Hour. In Education Technology and Computer Science (ETCS), 2010 Second International Workshop on (Vol. 1, pp. 606-609). IEEE.

[4] Abdullah, D., Radzi, S. M., Jamaluddin, M. R., & Patah, M. O. R. A. (2010, June). Hotel web site evaluation and business travelers' preferences. In Education Technology and Computer (ICETC), 2010 2nd International Conference on (Vol. 3, pp. V3-485). IEEE.

[5] Merono-Cerdan, A. L., & Soto-Acosta, P. (2007). External web content and its influence on organizational performance. European Journal of Information Systems, 16(1), 66-80.

[6] Morosan, C., & Jeong, M. (2008). Users' perceptions of two types of hotel reservation Web sites. International Journal of Hospitality Management, 27(2), 284-292.

[7] Yahaya, N. A., Gin, G. P., & Choon, C. W. (2005, March). Developing innovative web information systems through the use of web data extraction technology. In Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on (Vol. 2, pp. 752-757). IEEE.

[8] Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. International Journal of Contemporary Hospitality Management, 14(6), 274-285.

[9] Baloglu, S., & Pekcan, Y. A. (2006). The website design and Internet site marketing practices of upscale and luxury hotels in Turkey. Tourism management, 27(1), 171-176.

[10] Poon, W. C., & Lock-Teng Low, K. (2005). Are travellers satisfied with Malaysian hotels?. International Journal of Contemporary Hospitality Management, 17(3), 217-227.

[11] Choi, T. Y., & Chu, R. (2000). Levels of satisfaction among Asian and Western travellers. International Journal of Quality & Reliability Management, 17(2), 116-132.

[12] Segall, R. S., & Zhang, Q. (2008). Web mining of hotel customer survey data. In Proceedings of the 12th Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2008), Orlando, FL, June.

[13] Li, T., & Ruan, D. (2007). An extended process model of knowledge discovery in database. Journal of Enterprise Information Management, 20(2), 169-177.

[14] Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009, August). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on (pp. 1-5). IEEE.

[15] Luo, Y., Le, J., & Chen, H. (2009, October). A privacy-preserving book recommendation model based on multi-agent. In Computer Science and Engineering, 2009. WCSE'09. Second International Workshop on (Vol. 2, pp. 323-327). IEEE.

[16] Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1-24.

[17] Zhao, X., Wang, L., Guo, X., & Law, R. (2015). The influence of online reviews to online hotel booking intentions. International Journal of Contemporary Hospitality Management, 27(6), 1343-1364.

[18] Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. International Journal of Hospitality Management, 44, 120-130.

[19] Krsak, B., & Kysela, K. (2016). The use of social media and Internet data-mining for the tourist industry. J Tourism Hospit, 5(197), 2167-0269.

[20] Law, R., Buhalis, D., & Cobanoglu, C. (2014). Progress on information and communication technologies in hospitality and tourism. International Journal of Contemporary Hospitality Management, 26(5), 727-750.

[21] Jabreel, M., Moreno, A., & Huertas, A. (2017). Do Local Residents and Visitors Express the Same Sentiments on Destinations Through Social Media?. In Information and Communication Technologies in Tourism 2017 (pp. 655-668). Springer, Cham.

[22] (Hao, J. S. C., & Har, C. O. S. (2014). A study of preferences of business female travelers on the selection of accommodation. Procedia-Social and Behavioral Sciences, 144, 176-186.)

[23] Hananto, A. (2016). Application of Text Mining to Extract Hotel Attributes and Construct Perceptual Map of Five Star Hotels from Online Review: Study of Jakarta and Singapore Five-Star Hotels. ASEAN Marketing Journal, 58-80.

[24] DİRSEHAN, T. TEXT MINING IN THE HOSPITALITY SECTOR TO EXTEND THE MOTIVATION THEORY.

[25] Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59, 467-483.

[26] Dirsehan, T. (2015). An application of text mining to capture and analyze eWOM: a pilot study on tourism sector. S. Rathore, & A. Panwar in: Capturing, Analyzing, and Managing Word-of-Mouth in the Digital Marketplace, 168-186.

[27] Saltz, J. S. (2015, October). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 2066-2071). IEEE.

[28] Cenamor, I., De La Rosa, T., & Fernández, F. (2014). IBACOP and IBACOP2 planner. IPC 2014 planner abstracts, 35-38.

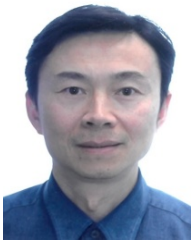
[29] van Eck, M. L., Lu, X., Leemans, S. J., & van der Aalst, W. M. (2015, June). PM^ 2: A Process Mining Project Methodology. In International Conference on Advanced Information Systems Engineering (pp. 297-313). Springer, Cham.

[30] Niaksu, O. (2015). CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing, 3(2), 92.

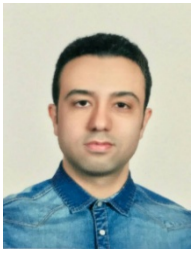
Authors

Mostafa Kamalpour received his Master degree in Information Technology Management from University Teknologi Malaysia in 2013, and granted his Bachelor degree in Business Information Technology from Limkokwing University of Creative Technology in 2011. His research areas are Data Mining, Information Systems, Tourism, and Social Media.

Atae Rezaei Aghdam is a lecturer of IT at Department of Computer Science at University of Applied Science and Technology, Tehran, Iran. He received his master degree in Information Technology Management from Universiti Teknologi Malaysia (UTM) in 2013 and a Bachelor of Software Engineering from Mazandaran Institute of Technology, Babol, Iran in 2012. His main research interests include Data mining in Tourism, Social media and Information systems.



Shuxiang Xu is currently a lecturer and PhD student supervisor within the School of Engineering and ICT, University of Tasmania, Australia. He received a Bachelor of Applied Mathematics from University of Electronic Science and Technology of China (1986), China, a Master of Applied Mathematics from Sichuan Normal University (1989), China, and a PhD in Computing from University of Western Sydney (2000), Australia. He received an Overseas Postgraduate Research Award from the Australian government in 1996 to research his Computing PhD. His current interests include the theory and applications of Machine Learning and Data Mining.



Ehsan Ghasem khani is a lecturer at Department of Computer Science, at University of Applied Science and Technology, Tehran, Iran. He received his Bachelor from Islamic Azad University of Arak and then granted his Master degree in Software Engineering from Islamic Azad University of Arak in 2009. His main research interests are Cloud Computing, Visualization, and Computer Network.

Aryan Baghi is currently studying Master Degree in Software Engineering at University of Science and Technology. He received his Bachelor degree from University of Science and Technology in 2016. His research interests are Big Data and Data Mining.