



DATA MINING FOR TRAVELS AND TOURISM

MANOJ B. KARATHIYA¹, RAVINDER SINGH SAKSHI², DILER SINGH SAKSHI³ AND DHAVAL R. KATHIRIYA⁴

¹M.C.A. Department, R. K. College of Eng. & Tech., India, Gujarat, Rajkot-360002.

²BCA, BScIT & MScIT Departments, Uttranchal College of Science & Tech., India, Uttarakhand, Srinagar-246174.

³B.C.A./PGDCA Departments, G.K. & C.K. Bosamia College, India, Gujarat, Rajkot-360005.

⁴Information Tech. Center, Anand Agricultural Uni., Gujarat, Anand-388110.

*Corresponding Author: Email- ¹karathiyamail@yahoo.co.in, ²ravindermca04@gmail.com, ³sakshidiler@gmail.com, ⁴drkathiya@gmail.com

Received: December 12, 2011; Accepted: January 15, 2012

Abstract- Data Mining may be described as the process of analyzing typically huge data sets to survey and find out earlier unidentified prototypes, styles and relationships to produce information for superior decision making. It has been fall out that today's aggressive situation for travel & tourism to raise their market and keep control these businesses would be pushed not to use data mining tools and techniques to develop, manage market tourism products and services. Objective of paper is to confer and demonstrate data mining and its application in travel and tourism.

Keywords- Data Mining, Predictive Modeling, Association Analysis.

Citation: Manoj B. Karathiya, Ravinder Singh Sakshi, Diler Singh Sakshi and Dhaval R. Kathiriya (2012) Data Mining for Travels and Tourism. Journal of Information and Operations Management ISSN: 0976-7754 & E-ISSN: 0976-7762, Volume 3, Issue 1, pp-114-118.

Copyright: Copyright©2012 Manoj B. Karathiya, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Now a day data mining can be measured a fairly freshly developed technology and methodology. It uses technologies of statistics, mathematics, machine learning and artificial intelligence. It aims to classify original, valid, useful, potentially and under stable correlations and patterns in data mining. Also it has been used intensively and comprehensively by marketers, financial institutes, retailers and manufactures. [1, 2, 3]

Data Mining Styles

Phase1: Pre Modeling

Initially a step in data mining is to find out the business difficulty. This is dangerous as one significant explanation for crash in data mining is the overemphasis on data analysis at the cost of the business problem to be addressed. On the other hand, stating a business problem by itself does not mechanically advise a data mining application. For business problems that are suitable for data mining, it is essential to explain the business problem into a data mining application. This is the second step in the pre modeling phase. Third step is to evaluate the data required and accessible for the data mining application. Data can arrive from in-house resource or outside resource. If the required data are not presented then they will have to be acquired or produced.

Fourth step in the pre modeling phase is also the most boring step for the preparation of the data mining. In various cases the necessary data are available but not in the same database or in the same standard format. Efforts will then have to be made to remove and join data from dissimilar database or sources and make them constant. More significantly accessible data may be partial or may contain errors. These data problems have to deal with too (filling lost data from other sources and correcting erroneous data).

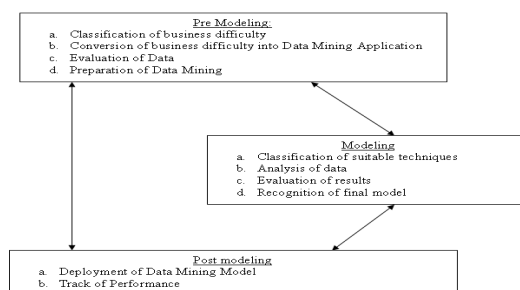


Fig. A- Data Mining Style

Phase 2: Modeling

Modeling phase can be judge to be the core of data mining. This is the phase where data analysis is performed. Normally, for any data

mining application, several data mining tools or techniques can be used. Therefore first step in the modeling phase is to recognize the suitable techniques to apply. Next step is to carry out the real analysis.

Behind the analysis, it is essential to review the result. This communicates to the purpose of the data mining application. Such as, if the objective is to do market segmentation, then it is necessary to review if the clustering results lead to interpretable, helpful and actionable market segments. Evaluation of results can be statistical in nature too. Data mining is an iterative process. Therefore, the evaluation may lead to a re-selection of the variables or are runs of the analysis etc.

Finally, if two or more models give acceptable results, then there is a need to identify the final model. Thus, the difference acceptable models can be compared with respect to their accuracy rates and the one that is most accurate can be selected as the final model.

Phase 3: Post Modeling

Post modeling phase relates to the events to be taken after the data analysis is finished. The first step is the deployment of the data mining model. What this involves depends on the objective of the data mining application. Such as, if the purpose is market segmentation, then the clustering results may be used for decision making.

Last step is the tracking of performance. This is essential because of changes in the environment in which an organization operates. Such changes may lead to a drop of the performance of data mining models, which may be dated. For instance the variables and relationship that help predict a target variable may change over time and a data mining model constructed in the past may no longer be useful at present. Therefore, tracking is important as deteriorating performance may signal the need to look at the data mining model again and to build an updated model if necessary.

As a final point, double headed arrows connecting the three phase of the data mining styles in Fig. A indicate that data mining is an interactive and iterative process. Very often, it is essential to travel back and forth among phases of steps when developing a data mining application. Such as, poor modeling results may mean looking at the data again. Therefore, data mining is a rigorously chronological process.

Data Tools

1. Description and Visualization

Description and visualization can contribute greatly towards understanding a data set and detecting unknown patterns in the data. As such, they are frequently performed before modeling is attempted in order to understand or delete relationship among variables. Description and visualization also help greatly in the summarization of data and in the presentation and reporting of results.

Description refers to the summarization of data to help understanding. An example of description is the profiling of data sets in order to understand their characteristics, similarities and differences. Standard description tools include summary statistics such as measures of central tendency, measures of dispersion and counts. Graphical approaches can also help to describe data and the relationship in data.

Visualization can be considered an enhanced graphical approach

that allows user input and interaction. An example is a rotating multidimensional plot that permits to define multiple dimensions in the plot as well as the direction and angle of rotation to facilitate viewing complex relationships. Colors can also enhance visualization.

2. Association and Clustering

In association, purpose is to determine variables which go jointly. It is a tool that looks for groupings or patterns among a set of items. Such as, analysis of market refers to a technique that generates probabilistic statements such as, if a tourist visits Nepal, there is a 0.66 probability that he/she also visit India. Such declarations (rules) are sensitive and easy to know. In addition, good association is expected to have predictive values. Though, many applications of association analysis are only exploratory in nature with a view to better understand grouping and patterns in the data set. Association rules can be useful for items building discount and promotion decisions, cross-selling etc.

Association analysis can be extended to include more sophisticated applications. For example, time sequence can be incorporated. To increase on line purchase, an online travel agency may try to isolate the sequence of web navigation that is likely or unlikely to lead to online purchase.

Clustering is an exploratory technique that attempts to discover natural grouping in data. The objective is to group similar objects into the same cluster and dissimilar objects into different clusters. Clustering is usually used to do market segmentation and to identify the cluster profile of the different segments. Knowing how the market is segmented and the characteristics of the different segments helps in decisions such as an organization's and the avenues and communications that can be used to reach the targeted segments.

3. Predictive Modeling [4]

Mainly ordinary and important applications in data mining generally involve predictive modeling, which can be further categorized into two main groups. Classification refers to the prediction of a target variable that is qualitative in nature. Estimation, on other hand refers to the prediction of a target variable that is quantitative in nature. Normally, predictive modeling attempts to predict a target variable on the basis of one or more input variables. Such as, multiple regressions can be used to predict the amount of tourist expenditure based on income, age and gender.

Neural network can also be used for predictive modeling. They can often model complex relationship in data well. Neural networks are modeled after the human brain, which can be perceived as a highly connected network of neurons.

Finally, decision trees can be used for predictive modeling too. They divide observations into naturally exclusive and exhaustive subgroups based on the levels of particular input variables that have the strongest association with the target variable. The end product can be graphically represented by a tree like structure, which is a compact explanation of the data. The end product can also be represented by explicit decision rules. Both representations are easy to interpret and use. Decision trees can model complex relationship reasonably well.

In practice, it is common to construct all the regression, neural

network and decision tree models and then assess the competing models to identify a final model.

Data Mining For Travel And Tourism

Travel and tourism industry is one of the main users of information technology [5]. Progresses information technology affects the services and facilities offered and how they are delivered and promoted. It's also affecting the organizational structure and interactions between customers and services providers [6]. Travelers are increasing used of internet and communication technology to find places that meet needs and expectation.

According to Buhalis integrated knowledge of tourist characteristics, images, attitudes and preferred destination attributes should be used to market destination more easily. By Magini hotels can use data mining to create direct mailing campaign, plan seasonal promotions, plan the timing and placement of advertisement campaigns, create personalized advertisement, define which market segments are growingly most rapidly and determine the no of rooms to reserve for wholesale customers and travelers [7].

Also there is data blast in the travel and tourism. Generation of centralized reservation and property management systems has resulted in big amount of data for hotels. At the same time, more access to more data [8, 9].

Association and Clustering

Association relates to the market basket analysis of hotels, airlines and other services among visitors for the principle of partner selection and marketing alliances. Market analysis of the preferred products among possible visitors is an important analysis to carry out before an investment is made [10].

By clustering segmenting possible travelers into different clusters based on personal information mined from their personal web sites. This would allow travel and tourism business for the possible understanding of travelers interest and needs then able to offer specially designed packages through email [11].

Predictive Modeling

Predictive modeling applications in travel and tourism industry consist of customer relationship management. Such as data mining can be carry out to classify probable visitor who are likely to reply direct mailing movement, visitors who under or over stay their reservation or type of services that visitors prefers. Also data mining can be used to predict possible value of each visitor, plan marketing views to hold visitors and produce information for visitor's correlation management [12].

Mining model would be used by the Department of Tourism of the Gujarat State to allocate suitable designed promotional material to selected visitors. Decision trees to understand visitor's preference and interactions with visitors in order to make and keep a loyal visitor relationship

Designs Of Data Mining Functions

1: Mounting Tour Packages

Top has observed an increasing number of tourists going to Gujarat from any city of India and wants to capture a greater share of this tourist segment. Presently, the schedules for Gujarat tours (especially Gujarat Tourism) cover a long list of the most common and popular tourist attractions. Current reply from best visitors has

indicated that not all these attractions request to all visitors. Subgroups of visitors seem to be interested in only particular subsets of attractions.

Some of best visitors who are eager to visit other cities of Gujarat and want to stay one or two days in Gujarat. To improve Gujarat tour packages, best wants to find out which tourist attraction in Gujarat can be grouped into subsets that would appeal of different subgroups of visitors. Best is considering offering a basic tour package (outside of Gujarat which is near) that incorporate special options to additional attraction in Gujarat and near to Gujarat for different subgroups of visitors.

Most popular places in Gujarat are: (a) Lothal Archaeological Site, (b) Ambaji Temple, (c) Naramada River, (d) Saputara Hill, (e) Aina Mahal, (f) Somnath Temple, (g) Gir National Park, (h) Kirti Mandir, (i) Dwarka, (j) Palitana Jain Temple.

To develop Gujarat tourism further, conducted a survey (through Blogger) of 1000 visitors who have visited Gujarat on tour packages of at least five days. As the survey is conducted through online Blog, online site and mail it was kept extremely simple by asking the respondents if they have greatly enjoyed visiting the popular attractions listed above to the level of wanting to visit again and to recommend to their relatives and friends. The positive response for an attraction is taken as an indication of the suitability of the attraction for inclusion in the Gujarat options. Data composed are then entered into SPSS (Statistical Package for the Social Science) file for analysis. SPSS is computer software used to do the data mining.

Result show that the five most popular attractions are: (a)Gir National Park (b)Lothal (c)Aina Mahal (Bhuj) (d)Saputara (e) Palitana Jain Temple. Additional analysis is performed using the SPSS Apriori Algorithm. In Table B show results.

Table A- Association Rules Can Be Constructed As Follows

Instances	Sup.Conf	Lift	Conseq.	Antecedent
507	50.7	91	1.794	Gir National Park Aina Mahal

Rules can be translated to mean that:

- a) Out of 1000 visitors surveyed, 507 would want to visit Aina Mahal again and would recommend Aina Mahal (Bhuj) to their relatives and friends.
- b) Out of these 507 visitors about 91 % would also want to visit Aina Mahal again and would suggested Gir National Park to their friends and relatives.
- c) Out of the 1000 visitors surveyed, 510 would want to visit Gir National Park again and would suggested Gir National Park to their friends and relatives. Out of 507 visitors mentioned in (b) 91% would do so. Therefore, life value is 91/51 or 1.784. Usually, a higher lift value indicates a more successful association rule.

Generally, the association rules suggest the following two Gujarat options: (a) Gir National park, Aina Mahal and Palitana Jain Temple and (b) Lothal and Saputara. These two options are expected to add the greatest value as a Gujarat expansion to the regular Gujarat tour packages inside of Gujarat [13].

Table B- Groupings Of Attraction In Gujarat

Fields	Percentage %	Count
Ambaji Temle	51.8	518
Gir National Park	51.5	515
Aina Mahal	50.4	504
Palitan Jain Temple	47.7	477
Lothal	46.3	463
Somanath Temple	45.2	452
Saputara Hill	42.5	425
Kirti Mandir	41.8	418
Dwarka	36.9	369
Narmada River	35.5	355

Table C- Association Analysis Result

In-stances	Supp.	Confidence	Lift	Consequent	Antec.
507	50.7	91	1.794	Gir National	Aina Mahal
511	51.1	90	1.761	Aina Mahal	Gir National
507	50.7	87	1.715	Jain Temle	Gir National
511	51.1	82	1.604	Jain Temple	Aina Mahal
515	51.5	73	1.417	Lothal	Saputara

2. Target Mailing Campaign

Best has performed a mailing operation a year ago to promote a new tour package to Gujarat. The advertising brochure was sent randomly to 2500 visitors and 831 visitors answered, giving a very good responses rate of 33.24%. The data the data related to this mailing campaign have been entered into a SPSS data set. Best is now considering a target mailing campaign for a parallel but new tour package to Gujarat. It considers that the relationship and patterns in the data can help it improve the response rate and reduce the cost of the campaign by targeting optional customers better.

The subsequent demographic characteristics are captures in the database: (a) Gender (b) Age (c) Education (d) Occupation (e) Response to the mailing campaign (f) Whether the visitors has taken any tour to Gujarat within Six month before the mailing campaign (g) Whether the visitors has taken any tour within Six month before mailing campaign other than Gujarat (h) Number of times the customer has traveled with the agency during the past three years.

The results show that visitors who have answered helpfully to the last mailing campaign for the Gujarat tour package are possible to be: (a) Visitors who are above 45 years old, female and who have taken a tour within six months before the mailing campaign to other than Gujarat or Saurashtra. (b) Visitors who are above 45 years old and who have neither taken a tour within six month before the mailing campaign to outside of Gujarat or Saurashtra nor to Gujarat and (c) Visitors who are above 45 years old, female and who have taken a tour within six months before the mailing campaign to Gujarat but not outside of Gujarat or Saurashtra.

On the other hand, the results show that visitors who have responded negative to the last mailing campaign for the Gujarat tour package are likely to be: (a) Visitors who are bellow 45 years and below. (b) Visitors who are above 45 years old and who have taken a tour within six months before the mailing campaign to Saurashtra as well as outside of Gujarat. (c) Visitors who are above 45 years old and male and who have taken a tour within six months before the mailing campaign to outside of Gujarat or Saurashtra but not to Saurashtra and (d) Visitors who are above 45 years old, male and who have taken a tour within six months before the mailing campaign to Gujarat but not to outside of Gujarat or Saurashtra.

Apart from for the variables appearing in the decision tree, the other input variables do not appear to be associated with the response to the last mailing campaign.

Decision tree results are expected to be useful to best in its next mailing campaign in targeting the right visitors. In this application, it is assumed that best has an up to date database of its customers with respect to the variables captured in the decision tree.

Termination

Data mining can be defined as the process of analyzing generally large data sets to search and find out previously unknown patterns, trends and relationships to generate information for better decision making. It is a very dominant and positive technology and methodology for travels and tourism industry, data mining can give very much to the ability of travel and tourism business to gain a competitive benefit and develop.

However, data mining is not without limitations. Some of majors are described as under. (1) Quality of data mining results and application depends on the availability and quality of data [14]. Also the data needed for data mining often exist in different setting and data systems. Hence, they have to be collected and integrated before data mining can be done. In addition problems such as missing data, corrupted data, inconsistent data etc. have to be determined before mining is completed. It has been expected that data preparation comprises about 75% of the resources needed for a data mining project.

(2) Mining of data may throw up patterns of some kind that are a product of random fluctuations [15]. This is especially so for large data sets with many variables. Data mining completed in a mechanical manner will not guarantee results or success. Always need human intervention, interpretation and judgment [7].

(3) Successful application of data mining needs the user to be knowledgeable in the domain area of application as well as in the data mining technology and tools. Domain knowledge is important to identify the appropriate business issues for develop data mining

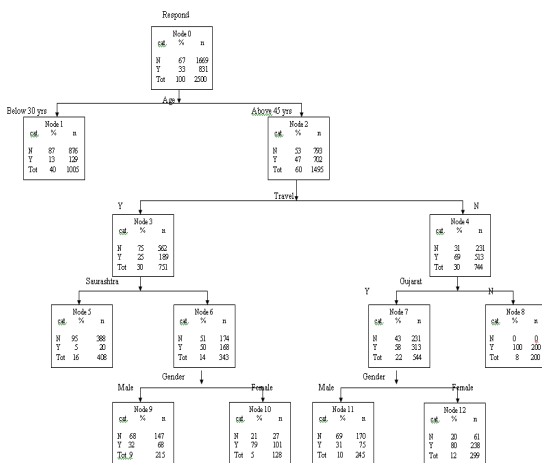


Fig. B- Decision Tree

application. It is also essential for specifying the appropriate models and correctly interpreting the results [16].

Finally, organizations developing data mining applications need to make a substantial investment of their resources in data mining. Data mining project can fail for a variety of reasons such as lack of management support and organizational commitment, unrealistic user expectations, poor project management, inadequate data mining etc [17].

Regardless of the limitation highlighted above, there is no doubt that data mining can play a critical role in travel and tourism research. What remains is for the travels and tourist industry and business to realize the potential benefits and usefulness of data mining.

References

- [1] Trybula W.J. (1997) *Annual Review of Information Science and Technology*, 32, 197-229.
- [2] Chung H.M. and Gray P. (1999) *Journal of Management Information Systems*, 16(1), 11-13.
- [3] Kreuze D. (2001) *Technology Review*, 104(2), 32.
- [4] Berry M.J.A. and Linoff G.S. (2000) *Mastering Data Mining: The Art and Science of Customer Relationship Management*. New York: John Wiley and Sons, Inc.
- [5] Sheldon P.J. (1997) *The Tourism Information Technology*. Wallingford: CAB International.
- [6] Olsen M. and Connolly D. (1999) *Tourism Analysis*, 4(1), 29-46.
- [7] Pyo S., Uysal M. and Chang H. (2002) *Journal of Travel Research*, 40(4), 396-403.
- [8] Magnini V.P., Honeycutt Jr. E.D. and Hodge S.K. (2003) *Cornell Hotel and Restaurant Administration Quarterly*, 44(2), 94-105.
- [9] Anonymous Hilton (1999) *Lodging Hospitality*, 55(13), 88.
- [10] Dev C.S., Klein S. and Fisher R.A. (1996) *Journal of Travel Research*, 35(1), 11-17.
- [11] Lau K.N., Lee K.H., Lam P.Y. and Ho Y. (2001) *Cornell Hotel and Restaurant Administration Quarterly*, 42(6), 55-62.
- [12] Kasavana M.L. and Knutson B.J. (1999) *Journal of Hospitality and Leisure Marketing*, 6(1), 83-86.
- [13] Min H.K., Min H.S. and Emam A. (2002) *International Journal of Contemporary Hospitality Management*, 14(6), 274-285.
- [14] Chopoorian J.A., Witherell R., Khalil O.E.M. and Ahmed M. (2001) *SAM Advanced Management Journal*, 66(2), 45-51.
- [15] Hand D.J. (1998) *The American Statistician*, 52(2), 112-118.
- [16] McQueen G. and Thorley S. (1999) *Financial Analysts Journal*, 55(2), 61-72.
- [17] Gillespie G. (2000) *Health Data Management*, 8(11), 40-52.