A Study on Data Mining Algorithms for Tourism Industry

Girish Kumar Sharma
Deptt. of Computer Applications , Bhai Parmanand Institute of PG Studies (Under DTTE)
GNCT Of Delhi

Promila Sharma
Deptt of Computer science , Mewar University , Rajasthan , India

Abstract: This paper presents different Data Mining Algorithms to be used to access data patterns in Tourism Industry. Data Mining , as the name implies is about mining data , to sieve through a lot of data to find useful information. Data Mining comprises techniques and algorithms for determining interesting patterns. Data Mining algorithms are becoming popular day by day in various applications like Travel & Tourism. The idea behind the data mining is very straight it is same like a human being become intelligent from examples and experience. In data mining; rules are developed by taking the behavior of given system (data set). Then these rules are used to evaluate the behavior / outcome for the given circumstances. The overall objective of this paper is to classify some well-known data mining algorithms.

Keywords: Travel & Tourism Industry , Data mining, Decision trees, Association Rules, Classification , Clustering.

I. INTRODUCTION

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data Mining comprises techniques and algorithms, for determining interesting patterns from large datasets. It applies machine intelligence and statistical tools to extract novel , useful and meaningful patterns in data which are not accessible through data query language.

Steps Of Knowledge Discovery (Data Mining)

This Knowledge Discovery in Databases (KDD) process consists of a sequence of the following steps

[1] Data cleaning – to remove noise and irrelevant data.
[2] Data integration – where multiple data sources are combined.
[3] Data selection – for retrieving from the database only the relevant data for the analysis.
[4] Data transformation – where data are transformed or consolidated into appropriate forms for mining.
[5] Data mining – the phase where the algorithms are applied in order to extract data patterns.
[6] Pattern evaluation – to find the interesting patterns which represents new knowledge.
[7] Knowledge presentation – when the visualization techniques are used to present the mined knowledge to the user.

What Data Mining Can Do

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

- Market segmentation - Identify the common characteristics of customers who buy the same products from your company.
- Customer churn - Predict which customers are likely to leave your company and go to a competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand what products or services are commonly purchased together e.g., Bread and butter.
- Trend analysis - Reveal the difference between a typical customer this month and last.

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 in-creasing used of internet and communication technology to find places that meet needs and expectation.

Tourism in India is the largest service industry, with a contribution of 6.23% to the national GDP and 8.78% to the total employment. The tourism industry in India generated about $100 bn USD in 2008 and that is expected to increase to $275.5 bn USD by 2018 at a 9.4% annual growth rate (India Tourism Statistics).

Today's aggressive situation for Indian travel & tourism industry is the need to raise their market and keep control their businesses to use data mining tools and techniques to develop, manage market tourism products and services. Modern technology has had a great impact to tourism (Domestic & International Both) since the 90s, when Internet technologies became the most dominant communication channel. Tourists' expectations are related not only to tourism services, but also to technology. [7]

In the tourism industry, knowing guests - where they are from, how much they spend, and when and on what they spend it- can help a company to formulate marketing strategies and maximize profits. Due to technological development, touristic companies have accumulated large amounts of customer data, which can be organized and integrated in databases that can be used to guide marketing decision [13]. Since identification of important variables and relationships located in these consumers -information systems could be a difficult task, some companies have attempted to raise the power of information by using data mining technologies. For example in hospitality area, the information systems have been used to assist the delivery of hospitality services. Some of key ways are [14] improved capacity management and operations efficiency, central room inventory control, last room available information, yielding management capability, marketing, sales and operational reports, tracking frequency flyers and repeat hotel guests, internal management of operations from transactions to human resources. Most of the items on the above list apply only to hotels and accommodation providers. In order to make high-quality marketing research and planning, data mining technology allows hotel companies to predict consumer behavior trends, which are potentially useful for marketing applications.

II. INTRODUCTION TO DATA MINING ALGORITHMS

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. To create a model, the algorithm first analyzes the data provided , looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.

Measuring Efficiency of an Algorithm

There are two aspects of algorithmic performance:
• Time
- Instructions take time.
- How fast does the algorithm perform?
- What affects its runtime?
• Space
- Data structures take space
- What kind of data structures can be used?
- How does choice of data structure affect the runtime?
In computer science, algorithmic efficiency are the properties of an algorithm which relate to the amount of resources used by the algorithm. An algorithm must be analyzed to determine its resource usage.

The Data mining model that an algorithm creates from your data can take various forms, including:

- A set of clusters that describe how the cases in a dataset are related.
- A decision tree that predicts an outcome, and describes how different criteria affect that outcome.
- A mathematical model that forecasts sales.
- A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together.

Data Mining Algorithms:
Several core techniques that are used in data mining describe the type of mining and data recovery operation.

- Classification algorithms predict one or more discrete variables, based on the other attributes in the dataset.
- Regression algorithms predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- Segmentation algorithms divide data into groups, or clusters, of items that have similar properties.
- Association algorithms find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- Sequence analysis
  algorithms summarize frequent
  sequences or episodes in data, such as a
  Web path flow.

2.1) Association Rules Algorithm

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.
Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional database. An association model consists of a series of item sets and the rules that describe how those items are grouped together within the cases. The rules that the algorithm identifies can be used to predict a customer's likely future purchases, based on the items that already exist in the customer's shopping cart for example:-
age(X,"20….29") ^ income(X,"20K…..29K") => buys(X,"CD Player")[support=2% , confidence=60%]
Where X is a variable representing customer. The Rule indicates that customers under study , 2% (support) are 20 to 29 years of age with an income of 20K to 29K and have purchased CD player. There is 60% probability (confidence or certainty) that a customer in this age and income will purchase a CD player.

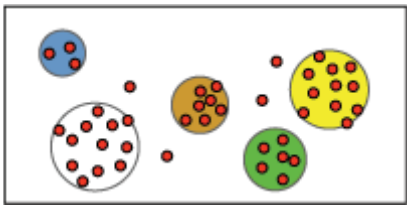
Association relates to the market basket analysis of hotels, airlines and other services among visitors for the principle of partner selection and marketing alliances. Market analysis of the preferred products among possible visitors is an important analysis to carry out before an investment is made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

2.2) Classification Algorithm

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model are derived based on analysis of a set of training data (The data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.
The derived model may be represented in various forms such as Classification Rules (i.e. IF…THEN rules) , Decision trees , Mathematical formulae or Neural Networks. Classification can be used to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, we can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, we might apply it into a particular class by comparing the attributes with our known definition. We can apply the same principles to customers, for example by classifying them by age and social group.
Classification predicts categorical (discrete , unordered) labels whereas Regression Analysis is a statistical methodology that is most often used for numeric prediction. Clustering allows to use common attributes in different classifications to identify clusters.

2.3) Clustering Algorithm

This algorithm uses iterative techniques to group cases in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions. This first identifies relationships in a dataset and generates a series of clusters based on those relationships. A scatter plot is a useful way to visually represent how the algorithm groups data, as shown in the following diagram. The scatter plot represents all the cases in the dataset, and each case is a point on the graph. The clusters group points on the graph and illustrate the relationships that the algorithm identifies.

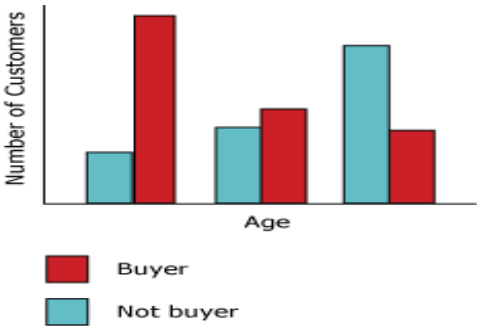


After first defining the clusters, the algorithm calculates how well the clusters represent groupings of the points, and then tries to redefine the groupings to create clusters that better represent the data.

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].

2.4) Decision Trees

Decision Trees algorithm is a classification and regression algorithm. For discrete attributes, the algorithm makes predictions based on the relationships between input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that you designate as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column.
The way that the algorithm builds a tree for a discrete predictable column can be demonstrated by using a histogram. The following diagram shows a histogram that plots a predictable column, Bike Buyers, against an input column, Age. The histogram shows that the age of a person helps distinguish whether that person will purchase a bicycle.



Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

2.5) Sequence analysis algorithms

Sequential Pattern Mining finds interesting sequential patterns among the large database. It finds out frequent subsequences as patterns from a sequence database. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining sequential patterns from their database. Sequential pattern mining is one of the most well-known methods and has broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. In the retailing business, sequential patterns can be mined from the transaction records of customers. For example, having bought a notebook, a customer comes back to buy a PDA and a WLAN card next time. The retailer can use such information for analyzing the behavior of the customers, to understand their interests, to satisfy their demands, and above all, to predict their needs.

III. RELATED WORK

Travel and tourism industry is one of the main users of information technology [5]. Progresses information technology affects the ser-vices 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]. Tourism has been regarded as one of the rapidly growing industries. The role of tourism in accelerating the economic development of a country has been widely recognized. It plays an important role in the economic, cultural, social and educational field and it is considered as the second largest economic activity in many countries for earning foreign exchange. Tourism is not a single industry but it is an aggregate of many components, capital investments in hotels , airways, roadways, railways, shopping centers, resorts and handicraft amounts to billions of dollars and millions of people earn their livelihood from direct and indirect employment in tourism industry. [12]

In the last decade, tourism has drawn much research attention in both the academic and industry community since the spending of tourists during their outbound/inbound traveling can benefit many service-oriented industries by creating many jobs and bring economic expansion or recovery for many countries or regions. [15]

With the fast growing of outbound tourism visitors in Asia Pacific regions in recent years, how to provide better quality services to these travel consumers is a crucial problem that the tourism industry should most concern. A new global trend is that more and more people rely on innovative e-Tourism IT solutions to search travel information or make reservations through search engines and travel agents online platforms. Hence, how to predict consumers traveling decisions based on their online behavioral or psychological patterns would be an interesting research topic. In this paper, a novel predictive model is proposed to find such consumer markers that can identify potential outbound visitors that can assist applications such e-Marketing of travel products, tourism e-Portal website development and outbound destinations promotion [4] , [16].

While discussing the overall challenges to the hospitality and tourism in India, the practitioners suggested measures which if addressed, could stimulate economic growth not only in this industry but also for India's economy as a

whole. The most significant factors affecting hospitality and tourism in India are: infrastructure management; government policy; workforce issues and education in hospitality and tourism; strategies for growth; crisis management; the management of destinations in India and the deployment of online techniques for marketing. [17]

To effectively extract information from a huge amount of data in databases , Data Mining Algorithms must be efficient and scalable. In other words the running time of a data mining algorithm must be predictable and acceptable in large databases. [1]

A variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. Many MNC's and large organizations are operated in different places of the different countries. Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions. [2]

The aim of Customer Relationship Management (CRM), the new management idea, is to improve the relationship between enterprise and the customer; its core is "understand customers, listen to customers", whose goal can be summarized as "attract potential customers, improve the existing customers' satisfaction and loyalty, and reduce customer churn". In one word, the ultimate goal of all is to improve profitability and competitiveness. As one of the three pillars, for tourism travel agencies to win from the market the importance of implementing CRM is self-evident. However, over a long period, the CRM data in the database is growing and people want to be able to provide a higher level of data analysis functions to convert data to be processed into useful information and knowledge automatically and intelligently. Data mining just provides us with effective method to solve these problems. It finds the potential relationship between data and provides decision support for us through the analysis of customer needs. [3]

Association Rule Mining finds application in market basket analysis. The market analysts would be interested in identifying frequently purchased items, so that the organization can adapt effective shelf space management and efficient sales strategies [10]. Two strategically measures of significant that control the process of association rule mining are support and confidence. Support is the statistical significance of a rule while confidence is the degree of certainty of the detective associations the entire process of association mining is controlled by user specified parameters, namely, minimum support and confidence. Many algorithms for generating association rules were presented over time. Some well known algorithms are Apriori and FP-Growth. Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of itemsets and uses a candidate generation function which exploits the downward closure property of support. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. FP-growth (frequent pattern growth) uses an extended prefix-tree (FP-tree) structure to store the database in a compressed form. FP-growth adopts a divide-and-conquer approach to decompose both the mining tasks and the databases. It uses a pattern fragment growth method to avoid the costly process of candidate generation and testing used by Apriori. [8]

Attractiveness of tourist destination is the research area of concern on tourism. It is defined as a pulling force generated by all the attractions of a specific area in a certain period According to the spatial extent of the study, different tourist attractions can be divided into different spatial scales. Attractiveness of tourist destination mainly refers to the mixture of tourist attractions which can meet the needs of the tourism market or means a tourist attraction group. It generally refers to the attractiveness of urban tourism [7]. Appraisal of tourism attraction and the tourism gravity model derived are important parts of theoretical research in tourism. Guiding the development of tourism activities, it affects the decision making, the methods, the spatial distribution and change direction of travel behaviors. Evaluation of the attractiveness of tourism destinations based on link analysis is based on calculating the link number of the official websites of the tourism attractions within the destination and analyzing the spatial distribution of the links. The approach is extended to the destination scale and a new travel attractiveness appraising method based on link analysis is proposed. The process includes four steps: (1) Ascertaining the range of a tourist destination.; (2) Confirming the tourism attractions in the range of the tourist destination. The common confirming methods include confirming by the local tourist guide to determine the attractions, by the determination of the local commercial companies and by the determination of the administration; (3) Calculating the travel attractiveness of each tourism attraction. Guided by the previous research conclusion the travel attractiveness of a tourism attraction can be indicated with the link account of its official website; (4) Confirming the travel attractiveness of the tourist destination by the travel attractiveness of each tourism attraction. [9]

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].

Web mining is the application of data mining on web data and web usage mining is an important component of web mining. The goal of web usage mining is to understand the behavior of web site users through the process of data mining of web access data.[18]

Due to technological development, touristic companies have accumulated large amounts of customer data, which can be organized and integrated in databases that can be used to guide marketing decision [13]. Buhalis (2002) said that the new generation travelers are more complex and highly demanding on quality of products. These travelers have known very well about attractions and tourism products. They had many experiences to spend time and money to travel. The new travelers like to compare details of products and choose the suitable items for themselves and they use the internet to search for information by themselves more than asking agency for services [14].

Recently, holiday tourism has developed rapidly and become a new economic growth-point of national economy. The rise of holiday tourism provides us new developing opportunity, yet, many problems remains due to the wideness and uncertain factors of holiday tourism. The realization of distributed sampling association rule mining algorithm in tourism was introduced for holiday tourism information data mining, improve a distributed sampling association rule mining algorithm DS-ARM, define the realization process of the algorithm and test the capability of the algorithm and use the algorithm in the analysis of the holiday traveler destination traveling behavior. [19]

Web usage mining is defined as a technique to mine the data in a database or web usage logs / records on the web. Web log is a list of reference data on web pages. Often associated as click stream data because each insert according to the lawyer-click the mouse button. This record can be analyzed from the perspective of both client and server. when evaluated from the perspective of the server, Sequential pattern is a collection of web pages ordered sequence that meets the support that has been determined and also the maximum . Support is not defined as a percentage of the session with a pattern, but rather the percentage of consumers who have a pattern. The most important application of web mining is targeted advertising. Sequential mining is the process of applying data mining techniques to a sequential database for the purposes of discovering the correlation relationships that exist among an ordered list of events. An important application of sequential mining techniques is web usage mining, for mining web log accesses, where the sequences of web page accesses made by different web users over a period of time, through a server, are recorded.[20]

IV. CONCLUSION AND FUTURE WORK

This paper has presented different algorithms of data mining. The overall goal is to evaluate the use of the data mining algorithms for Tourism Industry. Each algorithm has different goal and objective to classify the data set in different manner. In near future we will use different data mining algorithms to design the efficient algorithm for Tourism Industry to access Data Patterns The implementation of the given behavior will be done using some popular data mining tool like Weka, IBM SPSS , Advanced miner Professional etc.

REREFRENCES

[1] Jiwai Han , Micheline Kamber , *Data Mining and Techniques,* 2nd Edition , Simon Fraser University.

[2] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi.(2012).*The Survey of data mining Applications & Future Scope*, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[3] Shenglei , Pei.(2013) *Application of data mining technology in the tourism product's marketing CRM.* Instrumentation and Measurement , Sensor Network and Automation (IMSNA) , 2013 2nd International Symposium on DOI : 10.1109/IMSNA . 2013.6743300. IEEE Publications.

[4] Wu,E.H.C. , Jiang,B. , YaouHu.(2012). *Identifying Outbound Tourism Visitors by Using E-Services Behavioral and Psychological Markers* , Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on DOI: 10.1109/ICIS.2012.59 , Publication Year: 2012 , Page(s): 295 – 299

[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] Manoj B. Karathiya, Ravinder Singh Sakshi, Diler Singh Sakshi and Dhaval R. Kathiriya (2012) Data Mining For Travel and Tourism. Journal of information and Operations Management ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1.
[8] Vanitha ,K. , Santhi , R.(2011). *Evaluating the performance of association Rule Mining Algorithms* , Volume 2, No. 6, June 2011 , Department of Computer Studies, Saranathan College of Engineering , Trichy , India.
[9] LiliSong , LinaQi , JiQi , KunWang , Xiaojing Liu.(2010). *Evaluation of the attractiveness of tourism destinations based on link analysis* Geoinformatics, 2010 18th International IEEE Conference on DOI: 10.1109/GEOINFORMATICS.2010.5568078 , Page(s): 1 – 8
[10] Dev C.S., Klein S. and Fisher R.A. (1996) *Journal of Travel Re-search*, 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] Bartwal.(2008). *Incredible India Loosing Sheen* , McClatchy – Tribune Business News , Washington.
[13] Danubianu, M. and Hapenciuc, V. (2008). "Improving Customer Relationship Management In hotel industry by Data Mining Techniques", Proceeding of Competitiveness and Stability in the Knowledge-Based Economy, Vol: CD, 30- 31 Mai, 2008, Craiova, Romania, ISSN/ISBN: 978-606-510-162-3, Page: 2444-2452
[14] Buhalis, D. (2002). "eTourism: Information technologies for strategic tourism management". Financial Times Prentice Hall.
[15] World Travel & Tourism Council (WTTC), "The Impact of Travel & Tourism on Jobs and the Economy", 2003.
[16] Emel , G.G. , Akat.(2007). *Prolifing a Domestic Tourism Market by means of Association Rule Mining* , Anatolia.
[17] Jauhari (2009). *The Hospitality and Tourism Industry in India , Conclusions and Solutions.* Worldwide Hospitality and tourism Themes Vol 1 Issue , pp.75-80.
[18] Omar, R. , Md Tap, A.O. , Abdullah, Z.S.(2014). *Web usage mining: A review of recent works* Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International IEEE Conference on DOI: 10.1109/ICT4M.2014.7020638.
[19] DuJunping , ZuoMin , TuXuyana.(2008*). The realization of distributed sampling association rule mining algorithm in tourism* , Intelligent Control and Automation. WCICA 2008. 7thWorldCongresson DOI: 10.1109/WCICA.2008.4592921, IEEE Conference, Page(s): 183 – 187
[20] Gaol, F.L. *Exploring the Pattern of Habits of Users Using Web log Sequential Pattern* Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on DOI: 10.1109/ACT.2010.37 Publication Year: 2010 , Page(s): 161 - 163 IEEE PUBLICATIONS