Data Mining in Tourism

Indranil Bose
The University of Hong Kong, Hong Kong

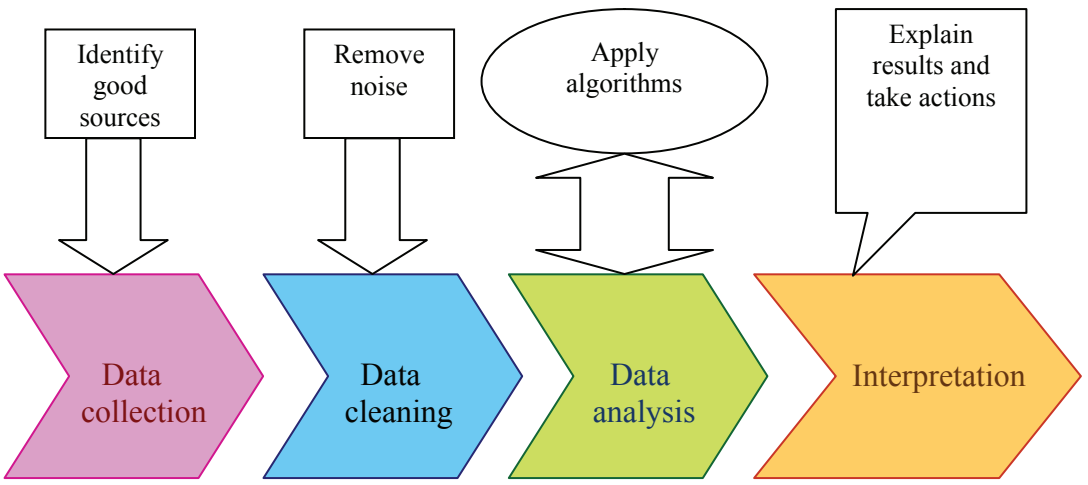
INTRODUCTION

Everyday, millions of people travel around the globe for business, vacations, sightseeing, or other reasons. An astronomical amount of money is spent on tickets, accommodations, food, transportation, and entertainment. According to World Travel and Tourism Council, travel and tourism represents approximately 11% of the worldwide gross domestic product (GDP) (Werthner & Ricci, 2004). Tourism is an information-based business where there are two types of information flow. One flow of information is from the providers to the consumers or tourists. This is information about goods that tourists consume such as tickets, hotel rooms, entertainments, and so forth. The other flow of information which follows a reverse direction consists of aggregate information about tourists to service providers. In this chapter we will discuss the second form of information flow about the behavior of tourists. When the aggregated data about the tourists is presented in the right way, analyzed by the correct algorithm, and put into the right hands, it could be translated into meaningful information for making vital decisions by tourism service providers to boost revenue and profits. Data mining can be a very useful tool for analyzing tourism-related data.

BACKGROUND

According to Tan, Steinbach, and Kumar (2006), "Data mining is the process of automatically discovering useful information in large data repositories" (p. 2). It uses machine learning and statistical and visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans. Data mining involves four key steps: (1) data collection, (2) data cleaning, (3) data analysis, and (4) interpretation and evaluation. During data collection the most suitable data need to be collected from the most appropriate sources. There is often a need to consolidate data from a number of sources. Cleaning or cleansing is the process of ensuring that all values in a data set are consistent and correctly recorded (Hui, Pandey, Steinbach, & Kumar, 2006). Obvious data errors are detected and corrected, and missing data is replaced in this step. The third and the most important stage of data mining is the analysis of the data using known techniques. Usually the analysis is done using statistical- or machine-learning-based approaches. The choice of the technique depends on the type of problem and also the availability of appropriate data mining software (Bose & Mahapatra, 2001). The most difficult part in data mining

Figure 1. Different steps in data mining



is interpretation of results obtained during data analysis. The analysis results may show a high degree of accuracy, but unless the accuracy can be related to the context of the data mining problem, it is of no use. Once some meaningful interpretation of the data analysis results can be done, the final step is to take action(s) so that the gathered knowledge can be put into practical use. In the case of tourism data mining, the data mining process goes through the same four steps. Figure 1 identifies the four steps involved in data mining and is similar to the one used in Bose and Pal (in press).

TOURISM DATA MINING

In this section we discuss the different types of machine learning techniques and explain how they have been used for analyzing data related to tourism. Usually two types of machine learning activities are common in tourism—association learning and classification learning. In association learning, the learning method searches for associations or relationships between features of tourist behavior. For example, the algorithm may try to find out if tourists who are interested in shopping also prefer to stay near the center of a city. That is, there is no specific target variable in this type of data mining, and so this is popularly known as unsupervised learning. A second style of machine learning is classification learning. This learning scheme takes a set of classified examples from which it discovers a way of classifying unseen examples. This is a form of supervised learning, in which there is a specific target variable**.** For example, by using classification analysts may be interested to classify tourists into two groups—high spenders and low spenders for luxury items. In this case the target variable is expenditure on luxury items. Based on a set of demographic and other variables the classification algorithm will establish the specific attributes of a tourist that qualify them as a high spender or a low spender. Next, we describe the various machine learning techniques used in tourism data mining.

Artificial neural networks (ANN). ANNs are nonlinear predictive models that learn through training (Jain, Mao, & Mohinuddin, 1996). They are composed of interconnected neurons. Each neuron receives a set of inputs. Each input is multiplied by a weight. The sum of all weighted inputs determines the activation level. A very powerful algorithm that is used in training ANNs is called *backpropagation*. Here, weights of the connection are iteratively adjusted to minimize the error based on the difference between desired and actual outputs.

Clustering. This is the process of dividing objects into groups whose members are similar in some way(s) (Han, Kamber, & Tung, 2001). Although there are many clustering algorithms, the commonly used ones are exclusive clustering and distance-based clustering. In an exclusive clustering algorithm if a certain datum belongs to a definite cluster then it cannot be included in another cluster. In distance-based clustering, if two or more objects are "close" according to a given distance they are grouped into the same cluster. Self-organizing feature map (SOFM) is a special type of an ANN-based algorithm that performs clustering.

Rough sets (RS). RS theory is proposed to address the problem of uncertainty and vagueness in the classification of objects (Slowinski & Vanderpooten, 2000). It is founded on the hypothesis that every object is associated with some information, and objects that are associated with the same information are similar and belong to the same class. The first step of RS is discretization of independent attributes where numeric attributes are converted to categorical attributes. The second step is formation of reducts that provide same quality of classification as the original set of attributes. The last step is classification of unknown data based on decision rules and reducts.

Support vector machines (SVM). SVM classifies an input vector into known output classes. It starts with several data points from two classes and obtains the optimal hyperplane that maximizes the separation of the two classes. For nonlinearly separable data, it uses the kernel method to transform the input space into a high dimensional feature space, where an optimal linearly separable hyperplane can be constructed. Examples of kernel functions are linear function, polynomial function, radial basis function, and sigmoid function (Chang & Lin, 2001).

Use of Data Mining in Tourism

Tourism policy makers, retail business executives, directors of scenic spot management companies, and government organizations want to know the relationship between tourism activities and preferences of tourists so that they can plan for required tourism infrastructures, such as accommodation sites and transportation. They also need detailed analysis to help them make operational, tactical, and strategic decisions. Examples of these include scheduling and staffing, preparing tour brochures, and investments. Due to the need for this analysis, formal statistical techniques were introduced in tourism. However, statistical techniques suffer from the drawback that several assumptions about distributions of data have to be made before any analysis can be conducted. If these assumptions are violated there is no guarantee that the results will be valid. This limitation of statistical methods has prompted researchers to use machine-learning-based data mining for tourism data analysis. The three main uses of data mining techniques in the tourism industry are: (1) forecasting expenditures of tourists, (2) analyzing profiles of tourists, and (3) forecasting number of tourist arrivals. In the following sections examples are presented to demonstrate how data mining techniques are used to support these activities.