ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2010. No. 3(99)
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Distributed Data Mining System for Tourism Industry

M. Danubianu
"Stefan cel Mare" University of Suceava, Faculty of Electrical Engineering and Computer Science
13 Universitatii str., Suceava, mobile: +40744.547164, e-mail:mdanub@eed.usv.ro

T. Socaciu
"Stefan cel Mare" University of Suceava, Faculty of Economic Science and Public Administration,
13 Universitatii str., Suceava, e-mail: socaciu@seap.usv.ro

D. Amariei
"Eftimie Murgu" University of Resita, Center of Advanced Research, Design and Technology,
1-4 Traian Vuia, 320085, Resita, Romania, e-mail: d.amariei@uem.ro

Introduction

Even in crisis times, tourism is a great income generator due to demand for its services that contains a wide range of transportation, accommodation, food and beverage, support services and travel distribution services. As a result, one of the strategic developments of the economy aimed the tourism industry. Strategies are based on different trends obtained from sophisticated analysis of data. Providing the managers in the tourism industry with information about and insight into the existing data is the key function of the data warehouse systems [1]. Data mining - techniques for exploration and analysis of large quantities of data in order to discover meaningful patterns and rules - helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, environment challenges. A system which enables the use of data mining techniques on data stored in a data warehouse is ideal for high quality analyses to support strategic decision. Designing and implementing a data warehouse is a complex and expansive process [2], so we can apply the data mining algorithms on large volumes of data from relational databases. The aim of this paper is to present the opportunity to use data mining methods on data from tourism and also to present two models of data mining systems, considering that the data is processed from a distributed database.

Information Systems and Applications Used in Tourism Industry

Information technology was initially viewed by the tourism industry as a back-office function that supports the finance and accounting areas. The industry has advanced far beyond this view during the past decade. As information is vital for tourism industry, effective use of information technology is necessary.

An information system regarding tourism activity should have some characteristics. It should collect, select and process information that is internal to the tourism activity coming from entities related to this sector such as National Tourism Agency and it should have subsystems that receive data from other business sectors, such as passengers landed from regular and nonregular flights, passengers flying with low-cost. It should support the decision-making process via integration of data from various sources, integrating them in a manner that permits analyses and comparisons between tourism indicators for the different regions of the countries and for different time periods. Also it should include historic information, such as the amounts of overnight stays per country of residence and for all types of hotel establishments throughout the 12 months of the year for every region [3]. It should also support the specialists who process and use tourism demand forecasts. As such, it detects the 10 countries with the highest tourism demand in the previous year. Taking into account the 10 countries listed and the forecasting methods suited to the specificity and nature of the data, the system prepares the necessary data to be processed by the specialists who create forecasts for tourism demand [4].

Currently, the most used information systems in Romanian tourism industry are the *front-office systems and the reservation systems*.

Front-office information systems are those data processing systems that provide reports in visual or written form. They are used mainly in the management of tourist accommodation (hotels, motels, hostels or cruise ships) or in the travel agencies activities. These systems may be used for: tourists registration when the personal data about tourists are collected; marketing of various tourism products, such as rental cars; rooms management, when are collected and processed data regarding the rooms status, (allows instant viewing of room availability for all room types, indicates whether rooms are dirty or clean, allows rooms to be placed out of inventory or out of order to restrict rental) and tracks of revenues, providing transaction processing and obtain information about any debts and credits in relation to customers

Information Systems Used for Reservations provide rapid access to information and ensures the accuracy of this information. They bring information services, booking and selling and are used both by individual tourists and travel agents or commissioners. Most often this type of systems uses Web technologies. These systems use hardware and software specific to conduct them activities. Although providers of tourist services in Romania currently use such systems for ticketing most, is well to remember that these systems can be used for marketing or management activities.

In the tourism industry knowing the 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 [5]. Since identification of important variables and relationships located in these consumer-information systems can be a difficult task, some companies have attempted to raise the power of information by using *data mining technologies*.

Data mining

Data mining is defined as the process of extracting interesting and previously unknown information from data, and it is widely accepted to be a single phase in a complex process known as knowledge discovery in databases (KDD). According to CRISP-DM [6], the reference model for this process, KDD consists of following phases:

- *business understanding.* This phase focuses on understanding project objectives and requirements from business perspective, on assessing of situation and determining the data mining goals.
- *data understanding.* Consist of raw data collection, describing data and verifying data quality.
- *data preparation.* Contains all activities needed to build the final data set from the initial raw data. Tasks include attribute selection and transformation, and cleaning data for modeling tools.
- *modeling.* Is the phase when are selected and applied various modeling techniques and the model is build.
- *evaluation.* Once build, the model is tested to be certain the proper model to achieve the project objectives and a decision regarding the use of the data mining results should be reached.
- *deployment.* Often the created model need to be presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be from simple generating of a report to complex implementing of a repeatable data mining process across the enterprise

In order to ensure that the extracted information generated by the data mining algorithms is useful, additional activities are required, like incorporating appropriate prior knowledge and proper interpretation of the data mining results. Figure 1 presents these phases and the most important interdepencies between them.

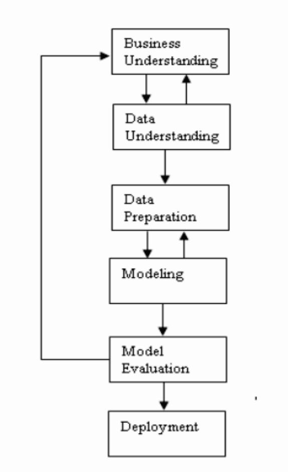

Fig. 1. Steps of the CRISP-DM process (adapted from [3])

Data Mining Techniques in Tourism Industry

One of the most important pieces of tourism industry is hospitality which is used to describe hotels and similar accommodations as well as restaurants and catering organizations and represent a very important aspect of the tourist industry. If hospitality organizations want to compete successfully, they must do so by using technology to drive value to both the customer and to the firm [7]. In this area information systems have been used to assist the delivery of hospitality services. Some of the key ways are [8]: improved capacity management and operations efficiency; central room inventory control; last room availability information; yield 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. The tasks performed by data mining can be grouped into the following five categories [9].

- classification arranges customers into pre-defined segments that allow the size and structure of market groups to be monitored. Also, predictive models can be built to classify activities. Classification uses the information contained in sets of predictor variables, such as demographic and lifestyle data, to assign customers to segments.
- clustering group customers based on domain knowledge and the database, but does not rely on predetermined group definitions. This function aids hoteliers to understand who their customers are. For example, clustering may reveal a subgroup within a predetermined segment with homogenous purchasing behavior that can be targeted effectively through a specific ad campaign with the scope that the members of the subgroup will increase their number of stays or become more loyal. On the other hand,

clustering may indicate that previously determined segments are not parsimonious and should be consolidated to increase advertising efficiency. Information such as demographic characteristics, lifestyle descriptors, and actual product purchases are typically used in clustering.

- deviation detection uncovers data anomalies, such as a sudden increase in purchases by a customer. Information of this type can prove useful if a hotel corporation wants to thank a guest for her or his recent increase in spending or offer a promotion in appreciation. Marketing managers may also attempt to draw correlations between surges in deviations with uncontrollable business-environment factors that are not represented in the database.
- association entails the detection of connections between records, driven by association and sequence discovery. For example, a possible association task could be employed in an effort to determine why a specific promotion was successful in one market, but ineffective elsewhere. Specific information regarding customer-purchase histories is necessary to formulate probabilistic rules pertaining to subsequent purchases.
- forecasting predicts the future value of continuous variables based on patterns and trends within the data. For instance, the forecasting function can be used to predict the future size of market segments.

Models for a Distributed Data Mining System

Data mining systems have the following characteristics: they must not limit the size of data sets, the performances are optimized for large data sets and they are enough flexible to use various techniques of data mining. Also they offer support for multi-user access and requires a total control over data access. Finally they provide management and maintenance at a distance.

The basic elements of a data mining system of data are: user interface, the specific data mining services, data access and data. Usually, data mining systems are built using client-server architecture, with different distribution on the two components of the items listed above. Fig. 2 presents architecture for data mining system.

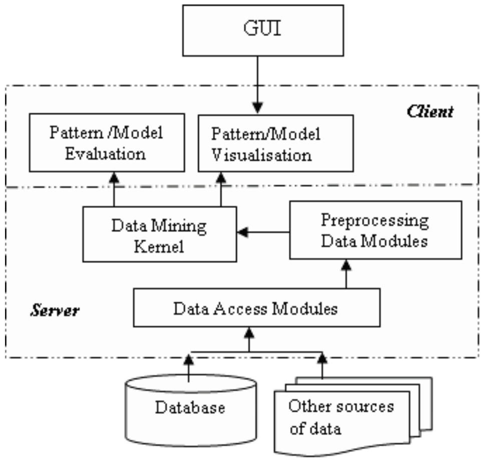

Fig. 2. Data Mining System Architecture

In order to achieve a prototype of a data mining system for hotel industry of Bucovina, we proposed two models of system architecture and we have studied some of their advantages and disadvantages. We started from the reality that each accommodation establishment manages its own data. Passing over specific needs all these work with databases containing data on customers, on services requested, on the amount spent, so on… If these systems allow a part of their data, in terms of ensuring data privacy, to be used for analysis, then it is possible that projections by necessary attributes of the tables to be available for sharing. If individual systems are connected through a communications network can assume that we are dealing with a heterogeneous distributed database, as shown in Fig. 3.

In order to study the two models we chose ten accommodation establishments that have agreed to share the data available for this purpose.

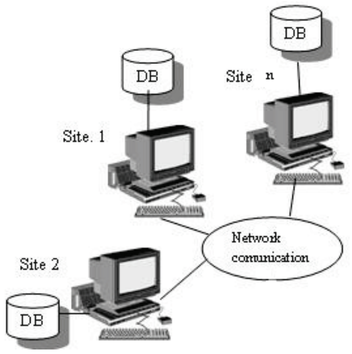

Fig. 3. Distributed Database Architecture

The first stage was a selection of required data. For that we have applied a projection by a list of fields with the same meanings of the tables, corresponding to the following model:

$$\prod_{c_{1k},c_{2k},\dots c_{nk}}(T_k), \tag{1}$$

where (c_{1k}, c_{2k}..c_{nk}) represent the list of n attributes required from table T_k.

The set of all these projections in a site form a fragment of the distributed database. Starting from this condition were analyzed the following situations. In the first case, were installed in each site, user interfaces and services suitable for data mining. Thus it was possible to apply local data mining methods (e.g. discovery of association rules). Local results were replicated in a single node and were combined to obtain the global solution. It is obvious that in the site where data were replicated were required additional operations in order to verify and validate the final results. For example, for association rules was necessary to calculate the global support and confidence. This approach has two major disadvantages. The first is related to the small volume of data processed on the local sites, which may lead to partial results inconclusive. The second disadvantage is the need to conduct further operations in the site where the results are collected.

The other option is a replication of all fragments in a single site where the specific components for data mining systems are installed. On these aggregated data we apply different methods of data mining. The advantage of this approach lies in that additional operations for further validation of the global results are eliminated. However, there is a drawback related to large volume of data transferred on the network.

We have made an experiment regarding the number of association rules obtained in the two cases, for different values of support. The results are presented in Fig. 4.

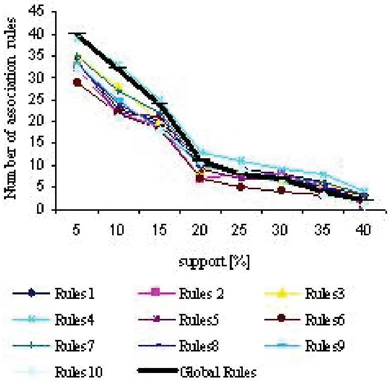

Fig. 4. The results of association rules detection for each of the ten sites and for global data set

We have used the Oracle Data Mining kernel for detecting the association rules. As we use similar features in all sites, it can be noticed that the number of rules is not different for the same support. Also, the figure shows that the global result is not dramatically affected if all data are replicated and are mined together.

Conclusion and Future Work

We have shown that data mining techniques can be applied successfully in the field of tourism, especially in connection with strategic marketing. We also kept in mind that these techniques can be applied to data sets from various sources, which can be successfully treated as fragments of a distributed database. In this context we have examined two models of distribution of components of a data mining system and we have underlined their advantages and disadvantages.

References

1. **Danubianu M., Socaciu T., Barila.** A Some Aspects of Data Warehousing in Tourism Industry // The Annals of the “Stefan cel Mare” University Suceava. Fascicle of The Faculty of Economics and Public Administration, 2009.
2. **Danubianu M.,** Advanced Information Technology – Support of Strategic Decision in Romanian Tourism Industry // 16-th International Economic Conference IECS2009, Sibiu, Romania, 2009.
3. **Poon A.** Tourism, Technology and Competitive Strategies. – Wallingford: CAB International, 1993.
4. **Ramos C., Perna F.** Information system for Tourism Activity Monitoring and Forecasting Indicators as an experience for Portugal // Tourism and Hospitality Research. – February, 2009.
5. **Danubianu M., Hapenciuc V.** Improving Customer Relationship Management in Hotel Industry by Data Mining Techniques // Competitiveness and Stability in the Knowledge-Based Economy. – CD. – 2008. – Craiova, Romania. – P. 2444–2452.
6. **Wirth R. and Hipp J.** CRISP-DM: Towards a standard process model for data mining // 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. – Manchester, UK. – 2000. – P. 29–39.
7. **Cassidy C., Chae B.** Consumer information use and misuse in electronic business: An alternative to privacy regulation // Information Systems Management. – 2006. – No. 23. – P. 75–87.
8. **Buhalis D.** eTourism: Information Technology for Strategic Tourism Management. New York: Prentice Hall, 2003.
9. **Magnini V., Honeycutt E. Jr., Hodge S.** Data Mining for Hotel Firms: Use and Limitations. – Cornell Hotel and Restaurant Administration Quarterly. – Sage Publication. – 2003.

Received 2009 11 13

M. Danubianu, T. Socaciu, D. Amariei. Distributed Data Mining System for Tourism Industry // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 3(99) – P. 31–34.

Romania has a huge tourist’s potential, but currently it is too little valued and exploited. As a result, one of the strategic developments of the economy aimed the tourism industry. The strategic decisions are based on different trends obtained from sophisticated analysis of data. Data mining helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, environment challenges. Two types of data mining systems are presented, considering that data are processed from distributed databases. Ill. 4, bibl. 9 (in English; summaries in English, Russian and Lithuanian).

М. Данубиану, Т. Сокациу, Д. Амарией. Распределенная система добычи данных для индустрии туризма // Электроника и электротехника. – Каунас: Технология, 2010. – № 3(99). – С. 31–34.

Румыния имеет огромный туристический потенциал, но в настоящее время он пока еще слишком мало ценится и мало эксплуатирован. Одна из стратегических разработок экономики направлена на индустрию туризма. На разных направлениях основанные стратегические решения получены применяя сложный анализ данных. Интеллектуальный анализ помогает предприятиям получить пользу из данных, которые с первого взгляда кажутся несвязанными. В результате принимаются конструктивные решения, на основе которых можно предвидеть, а не просто реагировать на проблемы окружающей среды. Цель заключается в представлении двух типов анализа данных системы, учитывая, что данные обрабатываются с распределенными базами данных. Ил. 4, библ. 9 (на английском языке; рефераты на английском, русском и литовском яз.).

M. Danubianu, T. Socaciu, D. Amariei. Paskirstytoji turizmo srities sistema „Data Mining” // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2010. – Nr. 3(99). – P. 31–34.

Rumunija turi didžiulį turizmo potencialą, tačiau jis menkai vertinamas ir eksploatuojamas. Todėl strateginės ekonomikos kryptys susijusios su turizmu. Strateginiai sprendimai priimami remiantis sudėtinga duomenų analize. „Data Mining” technologija padeda verslininkams reikšmingas sąsajas atrasti iš, atrodytų, niekaip nesusijusios informacijos. Pagal tai jie gali prognozuoti, o ne tik reaguoti į verslo aplinkos pokyčius. Pristatomi du „Data Mining” sistemų tipai, atsižvelgiant į tai, jog apdorojami iš paskirstytų duomenų bazių gauti duomenys. Il. 4, bibl. 9 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).