Explorative Data Analysis and Knowledge Modeling Methods for Marketing Decision Support Applied in the Tourist Sector

Stalidis, G.
Alexander T.E.I of Thessaloniki, Department of Marketing, Greece, 17th Klm Thessaloniki-Sindos, GR-57400

Corresponding author: E-mail: stalidgi@mkt.teithe.gr

Abstract: *The aim of this work is to support marketers in making informed decisions, such as the selection of target market segments, product positioning or optimal configuration of campaigns by extracting and managing knowledge from survey data. The proposed methods include the application of multivariate non-parametric factor analysis to reveal patterns underlying in survey data and their use as input to a knowledge engineering process, in order to build a knowledge-based system for marketing decision support in the tourist domain. To this end, an information technology framework is under development to extract from data and manage structured knowledge, thus making the results sharable, maintainable and available for problem solving to users who are not experts in analysis. In order to illustrate the methods, an analysis process has been performed on pilot survey data and a preliminary ontology-based knowledge model has been developed, that provides the necessary vocabulary to express the findings on tourist marketing. The results showed that the proposed methods are able to extract useful and reusable knowledge, applicable to marketing decision support in the tourist domain. Future work is planned towards the full development of a specialized Tourist Decision Support System.*
Keywords: *Marketing decision support, data analysis, knowledge modelling, tourist marketing, knowledge-based systems.*

1. INTRODUCTION

In this paper, work in progress and preliminary results are reported on the application of advanced data analysis and knowledge engineering technologies to support marketing decisions in the area of tourism. Considering that sources of market information are often limited or costly, it would be of great value for marketers to be able to make the most out of available survey data, to share and reuse information and, even further, to obtain and make use of the knowledge hidden in the data for solving specific problems. The focus of this work is to support marketers by extracting and managing knowledge from survey data. While in a conventional approach the analysis of primary survey data provides results in a static human-readable report, we are investigating methods that enable the capitalization of such results as reusable and expandable knowledge in electronic form. With the proposed scheme it will be possible to consolidate findings from many surveys in a single knowledge pool and to make them available for problem solving through computerized tools to users who are not experts in analysis. The proposed approach is to combine multivariate non-parametric analysis methods with rule-based knowledge management systems. More specifically, data analysis is based on the family of explorative factor analysis methods and in particular Multiple Correspondence Analysis (MCA) and Hierarchical Clustering (CAH) (Benzecri, 1992). The specific methods are suitable for revealing underlying patterns, such as customer profiles, market segments and product features that match specific demand and at the same time are suitable for relatively small datasets resulting from questionnaire-based surveys. The next step is to formalize the results for decision support. For this purpose, a knowledge engineering framework is being developed, consisting of a Knowledge Model (the structure and terminology used to express knowledge), a Knowledge Base (the operational component used to store and retrieve knowledge) and an inference engine (the mechanism that uses the stored knowledge to solve problems). An integrated data analysis and knowledge engineering platform is envisaged to support a range of marketing problems, including the optimal positioning of products based on the identification of trends, market segmentation and competition. The particular problem considered in the current work was to identify visitor expectations and travel patterns and to match them with the offerings and characteristics of hotels. Within the scope of this paper, the conceptual architecture of the sought framework is presented and a trial application of the data analysis/knowledge extraction methods has been performed on primary survey data. A preliminary ontology-based knowledge model has also been developed, that provides the base vocabulary to express the findings on tourist marketing.

2. BACKGROUND

2.1 Marketing decision support in tourism

The strong variability of the parameters that determine the success of tourism services make imperative the need for effective information-based marketing and a critical challenge is whether the tourist industry can respond to the evolution occurring in the composition and preferences of tourists, in order to maximize its competitiveness. There is also a clear shift to the use of information technology by both supply-side services and from the customer side (WTO, 2007). In other words, the way the tourist products are organized and promoted and the way customers search and choose their destinations are changing, as well as the expectations by different customer groups. Planning in this environment can be successful when based on qualitative information, advanced scientific methods and modern analytical tools (Gehrisch, 2005). From this perspective, the competitiveness of touristic destinations depends on their ability to develop, firstly, well-focused marketing and destination management strategies and secondly, effective control tools (Ritchie and Crouch, 2003). Strategic planning tools for improving the competitiveness of tourism in selected areas have been reported in the literature (Bouset07), based mainly on information management and less on sophisticated analysis and knowledge extraction. In another information-based approach (Wöber, 2003) the information needs and the corresponding sources have been defined for a Marketing Decision Support System in tourism. Knowledge-based decision support systems applied to tourism marketing have also been reported (Moutinho, 1996).

2.3 Knowledge engineering

One of the main goals of this work is to extract from a set of input data what is useful for decision support and to make it available for solving problems, in our case in tourist marketing. This process is usually referred to as Knowledge Extraction and can be based on Data Mining (or its synonym Knowledge Discovery in Databases-KDD) or Data Analysis and falls into the more general field of Knowledge Engineering (Baader, 2003). A large number of methods in this field have been reported in the last decades, stirred or facilitated by the advances in information technology, resulting in a large number of popular applications especially in business management and marketing (Shadbolt, 1999). By Knowledge Extraction we refer to any computerized process to create knowledge from sets of data or information. It is quoted that in Information Technology, Data is any set of numbers or text that can be captured from the real world, Information is data that are organised and linked to a particular meaning, while knowledge is sets of selected and properly formulated information that can be used to solve a particular problem. Finally, at the top level is Wisdom, which is the ability to generalize knowledge in order to solve unknown problems. What differentiates Knowledge Extraction from the abilities of a system based on Information is that it goes beyond offering structured information, to the creation of a model of the problem suitable for giving answers to a set of questions (Schreiber, 2008). Depending on the nature of the problem and the available data, Knowledge Extraction can be performed with a variety of methods, others based on statistical analysis, which are suitable to quantitative data, or to algorithmic approaches that handle logical relations among properties and are best suited to qualitative data (Han, 2001). The widely used term Data Mining refers to a set of methods suitable to extract knowledge in the form of hidden patterns from large volumes of data, typically maintained for other purposes, such as sales in the databases of transaction systems or website usage logs. There are several issues encountered in the design and development of a KBS (Ligeza06) and, as there are different types and usages of knowledge, an important challenge for any particular problem is to construct a suitable Knowledge Model that reflects the relevant to the problem part of the real world and is structured in a way that fits to the right techniques and tools to achieve the foreseen goals.

2.2 Data analysis as a knowledge extraction method

In order to effectively extract knowledge from primary survey data in any marketing or social survey, it is needed to couple the questionnaire with an analysis methodology that is suitable to qualitative data and is supported by a corresponding analysis tool. Considering the complexity of customer behavior and the dimensionality of problems related to social phenomena, it is clear that there is a need for an assessment method that is more sophisticated than conventional statistical analysis, which would be able to reflect a deep insight in the collected data (Van de Geer, 1993). Linear or other quantitative statistical models are not always sufficient, as they are based on the assumption of a specific model - that may not be known a-priori or its selection may be too restrictive - not to mention that conventional statistical methods are best suited to numerical data and in general do not perform well at problems involving nominal variables and decision logic (Roberts, 1996). On the other hand, widely used methods for knowledge extraction are methods in the category of Data Mining (Han, 2001) such as Mining of Association Rules, Classification by Decision Trees and On-Line Analytical Processing (OLAP), which are all popular techniques in Marketing for e.g. market segmentation and selecting audience for personalized promotion campaigns (Cooper, 2000). Such methods are suitable to the kind of data found in relational databases, which are data expressing properties of entities and relations between entities. In other words, data mining is good at handling categorical variables that describe interrelated concepts, seeking logical associations among specific values. However, there is an equally important drawback when the input data is available through surveys. Data mining is designed to run on large data volumes (e.g. the records of all company sales) and may produce unreliable results when applied to smaller datasets.

Considering the limitations of both statistical methods and data mining in extracting knowledge from questionnaire-based surveys, the approach investigated in this work is to apply methods from the branch of multidimensional factor analysis, namely Multiple Correspondence Analysis (MCA) and Hierarchical Classification (FACOR & VACOR) (Benzecri, 1992). These methods are effective in the analysis of qualitative characteristics, they start from the data itself and try by a progressive abstraction to discover patterns of which variables or group of properties are correlated (Greenacre, 2007). In this way it is possible to identify population clusters and highlight the properties which characterize more intensely each cluster (Stalidis, 2011). Moreover, multidimensional factor analysis has highly flexible data requirements and is therefore able to consolidate non-uniform data from a wide range of sources.

3. METHODS

3.1 Overall approach

The methods presented in this paper consist of two main parts, Data Analysis and Knowledge Modeling. Data Analysis includes the application of Multiple Correspondence analysis and Hierarchical Clustering to survey data in order to identify associations among properties and the Knowledge Modeling part supports the ability to handle them in a Knowledge-Based System (KBS). These two parts fit as important components of a wider framework for computerized knowledge management. In this paragraph, the overall approach for a tourist marketing decision support system is presented as the further goal to which the current work in progress is aimed at, while in the following sections we present the results obtained until now from a preliminary application.

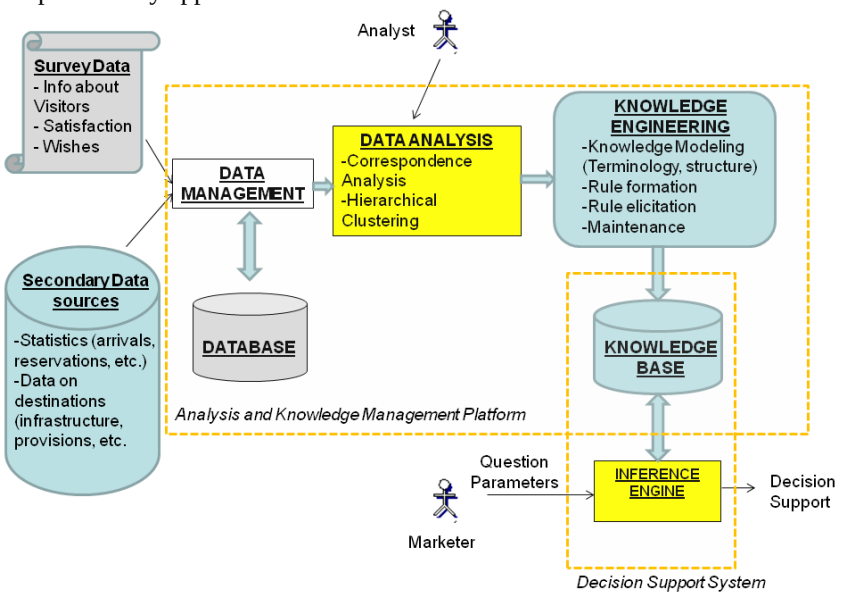

Figure 1: The overall architecture of the Data Analysis and Knowledge Engineering framework

The main components of the proposed architecture are:

1. **Data management**, including survey data collection and storage, quality control, importing of data from secondary sources, a data model defining the format of input data and data preprocessing.
2. **Data analysis**, which is used for knowledge extraction.
3. **Knowledge modeling**, which defines the knowledge structure, terminology and formalism. In our case, a rule-based model is adopted, matched to an ontology component. The ontology is used to provide the terminology necessary to express the rules and its creation is considered as a dynamic process that follows the data.
4. **Rule formation** is the process of transferring the analysis results to the rule syntax defined by the Knowledge Model.
5. **Rule elicitation** is performed by human experts and involves the selection of those rules that are useful for decision support and rejecting those that are naïve or not meaningful
6. **Knowledge Base** is the container of the produced knowledge and supports insertion/deletion, retrieval and maintenance tools, as well as consistency checking methods.
7. **Decision support** consists of the inference engine, which is able to use the accumulated knowledge to answer questions (i.e. check which rules apply and provide the conclusion) and also a user interface.

As shown in Figure 1, the analysis component is addressed to the user Analyst, who is responsible for exploring the input data. The Marketer is the user of the Decision Support system, he has access to the analysis results, but not to the original data. It is noted that data analysis is currently performed using the analysis software Méthodes d' Analyses des Données (MAD), which is a product of the Lab of Data Analysis and Multimedia Applications of the Department of Marketing,

ATEITH (MAD, 2012) and is used for teaching and research purposes (Karapistolis, 2002). It is planned that the core analysis components of MAD will be integrated with an open Knowledge-Based platform to construct a complete environment for survey analysis, knowledge management and decision support.

3.2 Application scenario and input data

In order to illustrate the method, an example usage scenario has been defined as a basis for trial application. The scenario is "data-driven" in the sense that it has been selected to match the data which were already available. This fits with a bottom-up approach, that is to search what we can make out of the data we can find, instead of firstly setting a concrete goal and then try to collect the necessary data. The application has been based on a primary questionnaire-based survey addressed to visitors, carried out within a student dissertation during the summer of 2010 at seaside destinations in Northern Greece. The questionnaire was designed to collect information about the expectations and satisfaction of the visitors regarding their hotel and also contained information about the characteristics of the visitors and the character of the visits. Given these data, the goal of the analysis was to find the preferences and requirements of visitors regarding their hotel, according to the characteristics of the visitors and the purpose of their visit and also to associate the degree of their satisfaction with the hotel characteristics which were the most significant for them.

3.3 Data Analysis

The sample consisted of 400 visitors and the questionnaire contained 19 questions of closed type. All questions were coded as categorical variables and the questions allowing multiple answers were coded into separate binary variables for each category. The resulting dataset consisted of 122 variables and 414 categories. Considering the large number of categories and in order to focus on more manageable individual problems, the dataset has been broken down to smaller (partially overlapping) subsets by selecting the variables which are desirable to associate in each particular exploration.

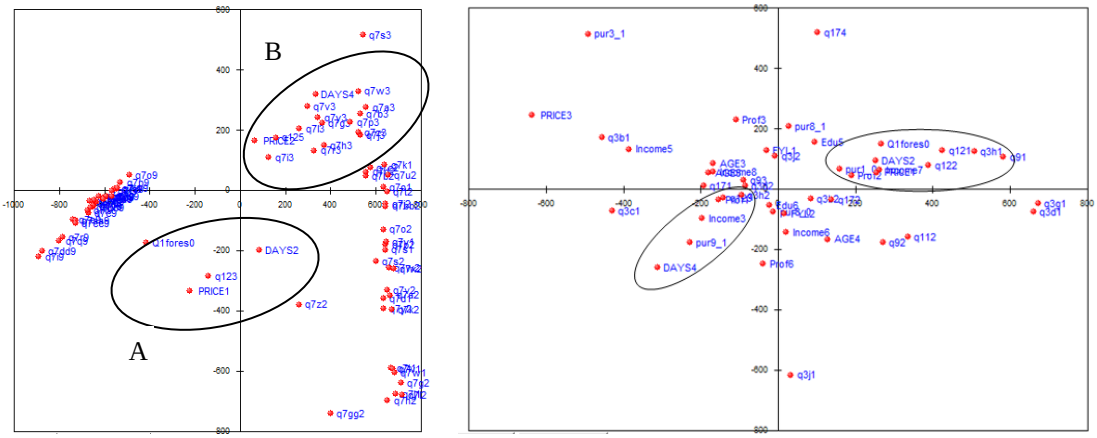

Figure 2: (a) The factorial plane 1X2 associating visit characteristics and expectations from the hotel and (b) the corresponding plane for the dataset associating characteristics of the visitor and type of visit

The first analysis included the variables related to the characteristics of the visit (e.g. reason for visit, duration, cost category), specific expectations from the hotel and a variable on the degree of general satisfaction. At this analysis step, parameters on visitor characteristics and details on satisfaction were not included. Multiple Correspondence Analysis (MCA) has initially been applied on the generalized contingency table (Burt). In Figure 2, the factorial plane 1X2 is displayed, which is formed by the first 2 factorial axes that interpret the largest amount of variance. On this plane (explaining 67% of the total variance) an interesting group was found, (marked as B in Figure 2) which shows that high general satisfaction, long visit duration (> 2 weeks) and low to medium price per night are associated with high expectations regarding special diet menus, entertainment activities, traditional hotel style and convenient location. In order to associate visitors' expectations with the purpose of the visit, a Hierarchical Clustering method (CHA) has then been applied on the same contingency table, clustering all categories into homogeneous classes. Indicative findings were that the purpose of visit "medical reasons" is associated with duration up to 4 days and low expectations for room service, internet, bar at the hotel, traditional style and location, whereas when the purpose of the visit is either a conference, sports or religious visit, there is intention for paying high or very high prices and the main expectations in these cases are facilities for people with special needs, athletic facilities, menus for special diet, facilities for relaxation and for the hotel to be part of a group of hotels. A similar analysis has then be applied, focused on the association between the characteristics of the visitor (e.g. age, profession, country of origin, etc.) and the type of visit. Interesting patterns identified on the factorial plane 1X2 were that (a) People working as private employees with high income (3000-3500 per month) that visited this destination for the first time, paid medium to low price per night (50-100€), the duration of their trip was 5-8 days and their source of information was a touristic agency, were dissatisfied from their hotel and would not recommend it. Finally, the analysis has been applied

on the variables related to satisfaction per specific point (facilities, personnel, location, etc.), associated with the corresponding expectations per point. One of the findings was that visitors who would definitely recommend their hotel in the medium-low price area and expected from the hotel to be suitable for families also wanted a convenient location and local cuisine and expressed their satisfaction mainly regarding the existence of swimming pool for children, sports facilities, special diet menus, recreation activities and comfortable hotel lobby.

3.4 Knowledge Modeling

In order to represent analysis results such as the ones presented above, it is required to (a) establish a terminology regarding the basic concepts (visitor, destination, transport, etc.) and their properties (age, country, etc.) in a hierarchical structure (e.g. *hotel* and *camping* are both *accommodation*), (b) express relations between concepts (e.g. a *visitor* **has requirement** *swimming pool*), (c) express more complex associations among properties in the form of rules. The proposed Knowledge Model to fit with these requirements consists of two components, an ontology to express terminology and structural relations and a rule-based component to express association rules. An ontology (Gruber, 1993) provides a description of the domain of interest, providing a common semantic base. This allows the integration of heterogeneous data sources and, more importantly, a formal basis which is the prerequisite for formal rule statement creation and inferential analysis. In a recent review of existing ontologies in the tourism sector (Prantner, 2007) it is found that these is a considerable number of efforts, such as the QUALL-ME (Ou, 2008) and the DERI e-tourism (DERI, 2011) ontologies. However, it is also found that all of them are problem-specific and it is unlikely that one of them will match the exact needs of the current problem. In this work we adopt a modular approach which involves the importing of suitable available ontologies as modules and complement them with a problem-specific dynamically created module. While this process is in progress, the current work is based on a preliminary ontology created to fit the specific application scenario. The implementation has been done on the **Protégé-OWL** platform (Protégé, 2012) using the Web Ontology Language (OWL), which is a family of knowledge representation languages for authoring ontologies that can be shared through the Web (OWL, 2012). Protégé is a free, open source ontology editor and knowledge-base framework, supported by a large community of developers and users.

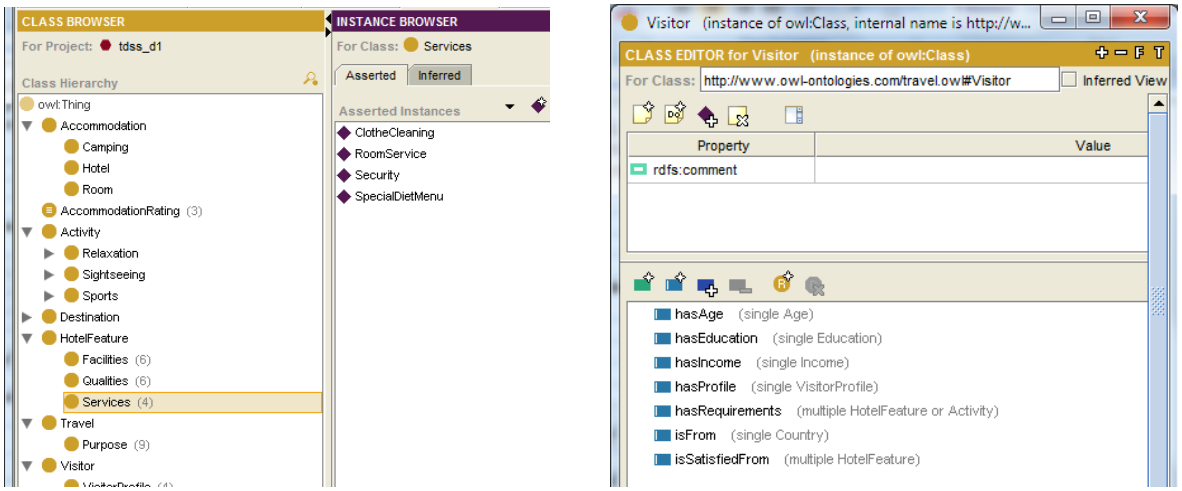

Figure 3: (a) The class hierarchy of the ontology in the Protégé environment and (b) the properties defined for class *Visitor*

Figure 3(a) illustrates the ontology class tree which represents our trial scenario. The classes and their subclasses are defined to reflect the basic concepts found in the tourist marketing modeling problem at hand, for example *Accommodation* is a general class representing all types of accommodation, while *Camping*, *Hotel* and *Room* are subclasses corresponding to more specific types. *Services* is a subclass of *HotelFeatures* and as *Services* we have defined the instances (that is specific objects in the class) ClotheCleaning, RoomService, Security and SpecialDietMenu. In Figure 3(b) the properties defined for the class *Visitor* are shown. Some Properties express a characteristic of Visitor, which takes values defined in another class (e.g. the property hasAge may take as value the age category 30-40, which is defined in the class Age). Other Properties express the relation of Visitor with other objects, e.g. the Property hasRequirements that links a Visitor with a *HotelFeature* or an *Activity*.

The rule-based (Ligeza 2006) component includes rules of the general form *C1* AND *C2* AND ... AND *Cn* → *E*, where *C1*, *C2*, ..., *Cn* constitute the conditions of the rule and E is the conclusion, action or decision. Each condition involves a property of a class and a comparison with a specific value, where the possible values should be defined as objects in the ontology. For example, a finding of the Data Analysis was that high general satisfaction, long visit duration (>2 weeks) and low to medium price per night are associated with high expectations regarding special menus and traditional hotel style. This is expressed as a rule using the vocabulary defined in the ontology as:

IF *visit* hasDuration Long **AND** *Hotel* hasPrice LowMedium **THEN** *Visitor* hasRequirement SpecialDietMenu, Visitor hasRequirement TraditionalStyle

In a similar way, all the analysis results can be formalized and introduced in the Knowledge Base. The Protégé framework offers tools for exporting to various formats, publishing on the Web and importing external ontologies that can be synchronized with the existing one. These capabilities facilitate the ability to reuse and share the knowledge content. The next step is to connect to an inference engine in order to execute queries or, in more simple words, to use a program that allows a user to ask questions that the engine answers on the basis of the stored knowledge.

4. CONCLUSIONS

In this paper, a method is illustrated for analyzing marketing survey data to reveal interesting patterns for marketing and for expressing the results in a structured, computerized form, to be available for knowledge-based decision support systems. In this way, results from many different surveys, as well as secondary information already available can be consolidated in a common knowledge pool that is sharable and usable by non-experts in analysis. The application on a trial scenario in tourism marketing showed that the proposed methods are able to extract and manage useful knowledge, applicable to marketing decision support in the tourist domain. The advantages of the proposed analysis methods compared to common statistical analysis and data mining are that they are suitable to the smaller datasets and qualitative data found in questionnaire-based surveys. Future work is planned towards the development of a specialized Tourist Decision Support platform.

Acknowledgment

This work is supported by the Research Program Archimedes III, project "Data Analysis and Knowledge Management Technologies for Planning Tourism Products". This work was conducted using the Protégé resource, which is supported by grant LM007885 from the United States National Library of Medicine.

References

Baader, F., Calvanese, D, McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Benzecri, J.-P, (1992), *Correspondence Analysis Handbook*. New-York: Dekker, P.

Bousset, J.P., et al, (2007). 'A Decision Support System for Integrated Tourism Development: Rethinking Tourism Policies and Management Strategies',*Tourism Geographies*,9:4,387 — 404.

Cooper L.G., Giuffrida G. (2000), "Turning Datamining into a Management Science Tool: New algorithms and Empirical Results", Management Science, vol 46, No 2, February 2000 pp. 249-264.

DERI, OnTour Ontology, http://etourism.deri.at/ont/index.html, last visited: 15/12/2011.

Gehrisch, M. (2005). "What is a CVB". in R. Harrill, Fundamentals of Destination Management and Marketing. Educational Institute of the American Hotel & Lodging Association. Michigan, U.S.A.

Greenacre, M., 2007, Correspondence Analysis in Practice, Chapman & Hall.

Gruber,T.,(1993), "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, 5(2), pp. 199-220.

Han J. and Kamber M., (2001), *Data Mining – Concepts and Techniques*, Academic Press, San Diego.

Karapistolis, 2002, "The MAD software", Data Analysis Bulletin, vol 2, pp 133 (in Greek).

Ligêza, A., (2006), Logical foundations for rule-based systems, Studies in Computational Intelligence, vol. 11, 2nd Ed., Springer-Verlag Berlin Heidelberg.

Méthodes d' Analyses des Données - MAD (2012), http://www.mkt.teithe.gr/index.php/el/ereuna/65/202-mad , last visited: 30/4/2012.

Moutinho, L., Rita, P. and Curry, B. (1996). Expert Systems in Tourism Marketing, London and New York: Routledge.

Ou, S., Pekar, V., Orasan, C., Spurk, C., Negri M., (2008), Development and Alignment of a Domain-Specific Ontology for Question Answering, in European Language Resources Association (ELRA) (ed.): Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

OWL, W3C Recommendation (2012), http://www.w3.org/TR/owl-features/ , last visited: 30/4/2012.

Prantner, K., Ding, Y., Luger, M., Yan, Z. (2007), "Tourism Ontology and Semantic Management System: State-of-the-Arts Analysis", Proceedings of IADIS International Conference WWW/Internet 2007, pp.111-115.

Protégé, (2012), http://protege.stanford.edu/ , last visited:30/4/2012.

Ritchie, J.R.B. G.I. Crouch (2003). The Competitive Destination: A Sustainable Tourism Perspective. CABI Publishing, Wallingford.

Schreiber, G. (2008). "Knowledge Engineering", in: *Handbook of Knowledge Representation*, F. van Harmelen, V. Lifschitz, B. Porter (Eds.), Elsevier, 2008, pp. 929–946.

Shadbolt N. and N. Milton, (1999). "From knowledge engineering to knowledge management", *British Journal of Management*, vol. 10, 1999, 309–322.

Stalidis, G. and Karapistolis, D. (2011), "Data Analysis to support business planning", 6th Panhellenic Data Analysis conf with international participation, Thessaloniki.

Van de Geer, J. P. (1993). Multivariate Analysis of Categorical Data: Applications Newbury Park: Sage Publications Inc. Advanced Quantitative Techniques in the Social Sciences Series Vol 3.

Wöber, K.W. (2003) Information supply in tourism management by marketing decision support systems. Tourism Management, 24 (3), 241–255.

World Tourism Organization/WTO (2007). *A practical guide to Tourism Destination Management*. Madrid, Spain.