Manuscript received Jul. 10, 2008; revised Sep. 17, 2008

Application of Business Intelligence methods for personalizing tourist services

ALBERTO SALGUERO, FRANCISCO ARAQUE, CECILIA DELGADO
Department of Software Engineering :: ETSIIT
University of Granada (Andalucía)
C/ Periodista Daniel Saucedo Aranda s/n
SPAIN
agsh, faraque, cdelgado {@ugr.es}

Abstract: The tourist web sites usually provide static information which can be accessed by mean of some kind of search/filter forms. The next step in the evolution of this services is to include some kind of Decision Support System which easy the task of finding what the user is looking for. In this paper we describe a system capable of generating the routes which better fits the user characteristics and preferences. The system uses a clustering/classification process for selecting the most interesting locations to visit according to the list of most historically visited locations included by similar tourist in their travels. Once determined the points of interest the system performs a process for selecting the most adequate routes connecting those points, taking into account the tourist characteristics and the knowledge about the location to visit.

Keywords: Tourism, personalization, data warehouse, clustering, classification.

1 Introduction

Planning a trip is a common but complicated human activity which requires some reasoning capabilities [2]. It is not easy to make automatic tools which encompass the whole planning process. It is usual that, when someone wants to plan a trip to certain place which he has not visit before, he searches for interesting points of interest (POI) in that place for visiting. With the grown of Internet and the generalization of Web 2.0 there are a lot of information that should be read prior of selecting the most interesting spots [16], [20]. Once selected it is necessary to search for information about where they are and how to reach them. Usually, when the area to visit is wide enough it is advisable to take the public transport. This implies to search the Web for finding the bus/metro lines…

According to [10], the tourism in Spain has considerably contributed to its economy, representing the 10'8 % (106.374 millions of Euros) of the GDP of the country. The Spanish government, concerned with this issue has incentivized the development of projects which aims to raise the quality of the touristic services. Undoubtedly and according to some analysis performed by the government, one the most fundamental aspects in the period of time between the 2004 and 2007 is the grown of Internet. One of these studies states that the people who seek information about their trip in Internet represent the 51'6 % of the total [9]. Given this situation the Spanish government has impulse the projects relying on this technology, giving as result, for instance, the touristic information systems presented in figure 1 [6], [5]. However, E-Commerce web sites in general and tourist Web portals in particular should provide more functionalities than a conventional document-based web search [15]. These web sites should incorporate some kind of decision support functionalities to assist users in selecting the most appropriate content [13].

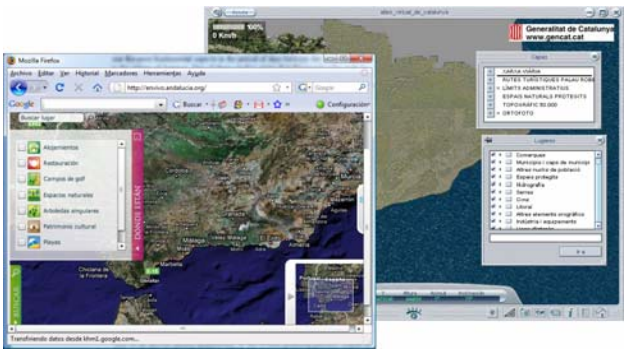

Fig. 1. Examples of touristic information systems developed by the Spanish government.

There are certain tools in which might help us to plan the trip, like Google Earth/map, GPS navigators, Microsoft MapPoint/Virtual Earth and other mapping tools. The problem of using these tools is that they always give the shortest path between locations. This is useful in most of cases but it is not when planning a trip. There are often more interesting routes than the

shortest. The problem is that usually these routes are only known by people familiar with the area to visit.

We have developed a system for solving these two issues: make easy the selection of the most interesting spots and planning the trip according to the characteristic of the customer. The system is based, as shown in the Fig. 2, in Data Mining techniques. In fact, we use a clustering process for finding the profiles of the customers and then we use this information for offering him the most interesting points of interest according to the historical data recorded by the system. Once the customer has selected the POIs he wants to visit the system is able to find the route between them which better fits to his personal characteristics (age, overall physical condition…) and to other environment conditions like weather forecast or timetables.

There are some interesting tools developed for facilitating the laborious task of planning a trip. In [2] the *Travel Assistant* tool is presented. It provides an interactive approach to making travel plans where all of the information required to make the trip is presented to the user in form of choices. The election of an answer affects dynamically to the overall planning process. [1] introduces the concept of travelers' preferences and encourages the need of using intelligent techniques for adapting the route to them. They point to some approaches, like machine learning, fuzzy logic and heuristic searches, which may be used for adapting the routes to the traveller's characteristics. The system presented in [19] plans the trips step by step interactively querying the travel agent about the customer's preferences. Once the trip satisfies the customer's constraints the route is validated and presented to him.

In our case, we have based our work in a common technique usually used in the business area: a clustering/classification process for learning the customer profiles coupled with a huge data base containing historical information. We base our work in the assumption that similar customers are interested in the same things.

The remaining part of this paper is organized as follows. In Section 2, the concepts of DW and Data Mining are revised; in section 3 our architecture is presented; Section 4 explains in depth the process of route generation. In section 5 an example is presented in order to clarify the process presented in this work. Finally, Section 6 summarizes the conclusions of this paper.

2 Data Warehouses and Data Mining

It is obvious that there is no organization running without data. The data can be viewed as tangible assets of an organization just as any physical asset. So, they need to be stored and made available to those who need them in order to be used at any moment. Since the data by themselves are useless, they must be put together to produce useful information. In turn, information becomes the basis for relational decision making. To facilitate the decision-making process, a new piece of technology more sophisticated than a database system was developed and called Data Warehouse (DW). The DW can be generally described as a decision-support tool that collects its data from operational databases and various external sources, transforms them into information and makes that information available to

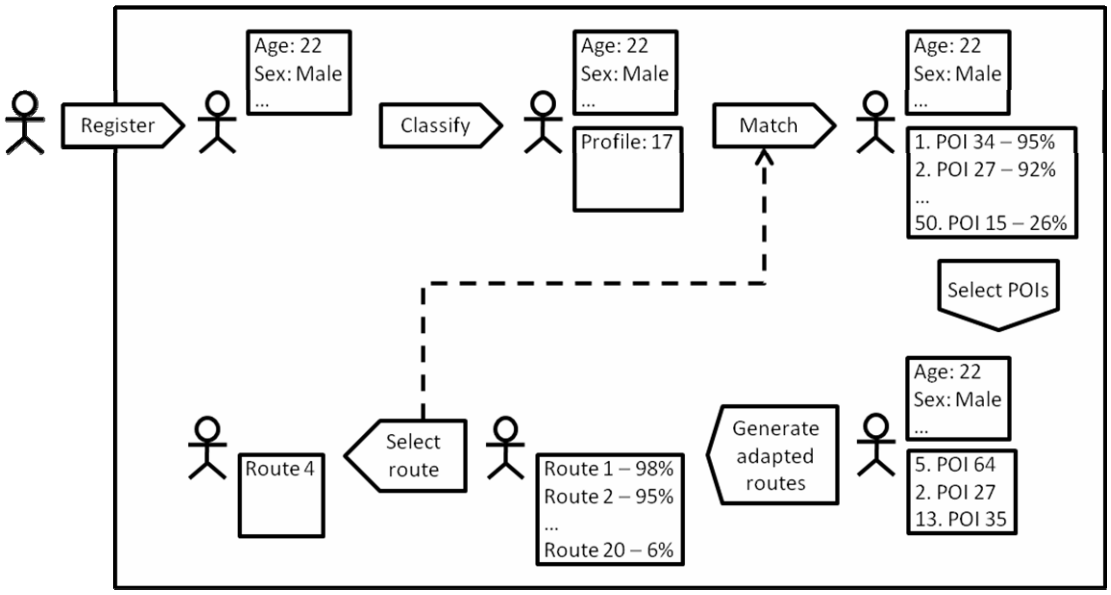

Fig. 2. General system workflow

decision-makers (top managers) in a consolidated and consistent manner [11][12]. The persistence of huge amounts of data (possibly distributed and heterogeneous) opens a new perspective for various statistical analysis methods which are essential for strategic decisions in tourism.

Inmon defined a DW as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process" [11]. A DW is a database that stores a copy of operational data which structure is optimized for query and analysis. The scope is one of the DW defining issues: it is the entire enterprise. Related to a more reduced scope, a new concept is defined: a data mart is a highly focused DW which scope is a single department or subject area. The DW and data marts are usually implemented using relational databases defining multidimensional structures. The generic architecture of a DW is illustrated in Fig. 3. Data sources include existing operational databases and flat files (i.e., spreadsheets or text files) in combination with external databases. The data are extracted from the sources and then loaded into the DW using various data loaders and ETL tools. ETL stands for extract, transform and load, the processes that enable companies to move data from multiple sources, reformat and cleanse it, and load it into another database or on another operational system to support a business process. The warehouse is then used to populate the various subject (or process) oriented data marts and OLAP servers. Data marts are subsets of a DW categorized according to functional areas depending on the domain (problem area being addressed) and OLAP servers are software tools that help a user to prepare data for analysis, query processing, reporting and data mining. Thus, a DW coupled with OLAP enables managers to creatively approach, analyze and understand the problems. The OLAP analyzes data using special DW schemas and enables users to view data using any combination of variables.

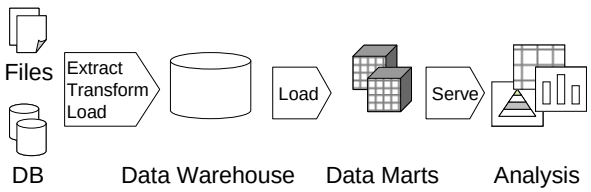

Fig. 3. Data Warehouse general architecture

We can define Data Mining (DM) as the process of extraction of interesting information from the data in databases. According to [7] a discovered knowledge is interesting when it is novel, potentially useful and non-trivial to compute. The use of Data Mining processes will help us to find out the patterns, features and in general the knowledge we are looking for. In fact to find out the features, patterns, etc. we have used a process of clustering and classification for this task.

The use of a clustering/classification process can be considered a DM method for finding homogeneous groups of individuals and acting according to the knowledge about the group [18]. The great quantity of data which can be extracted from a web-based application makes it a great and an easy tool for finding out the preferences of the users [14].

3 System architecture

The DW system coupled with OLAP functionalities is used to provide solutions for marketing problems, since it transforms operational data into strategic decision-making information. The DW stores summarized information instead of operational data. This summarized information is time-variant and provides effective answers to queries such as "Which kind of customers can we find?", "What is the profile of a specific client?" and so on.

The architecture of the system, as shown in Fig. 4, is a DW oriented architecture. It is based on the architecture presented in [17]. The DW is the central repository of data. It provides all the necessary information to all of the functional modules in the system and stores the resulting information. The main functionality of each module is:

- *Extract, transform and load*: This module is responsible of extracting and loading the customer personal information and the information the system needs to per-form the planning of the trip (public transport timetable, weather forecast…) in the DW. The customer personal information is easily loaded in the DW because it is obtained through the user registering process of the Web portal, i.e. we can modify it as needed. The rest of the information is gathered mainly from inde-pendent Web data sources. This means that we need to develop wrappers for ac-cessing to their information. We have developed some tools for extracting this information efficiently, polling the Web every certain time for detecting changes [4][3].
- *Clustering*: Given a list of users, this module is capable of performing a clustering process, obtaining the list of customer profiles as the result. These customer profiles correspond to the cluster centroids. This process is carried out every certain period of time. It is independent of the other processes.

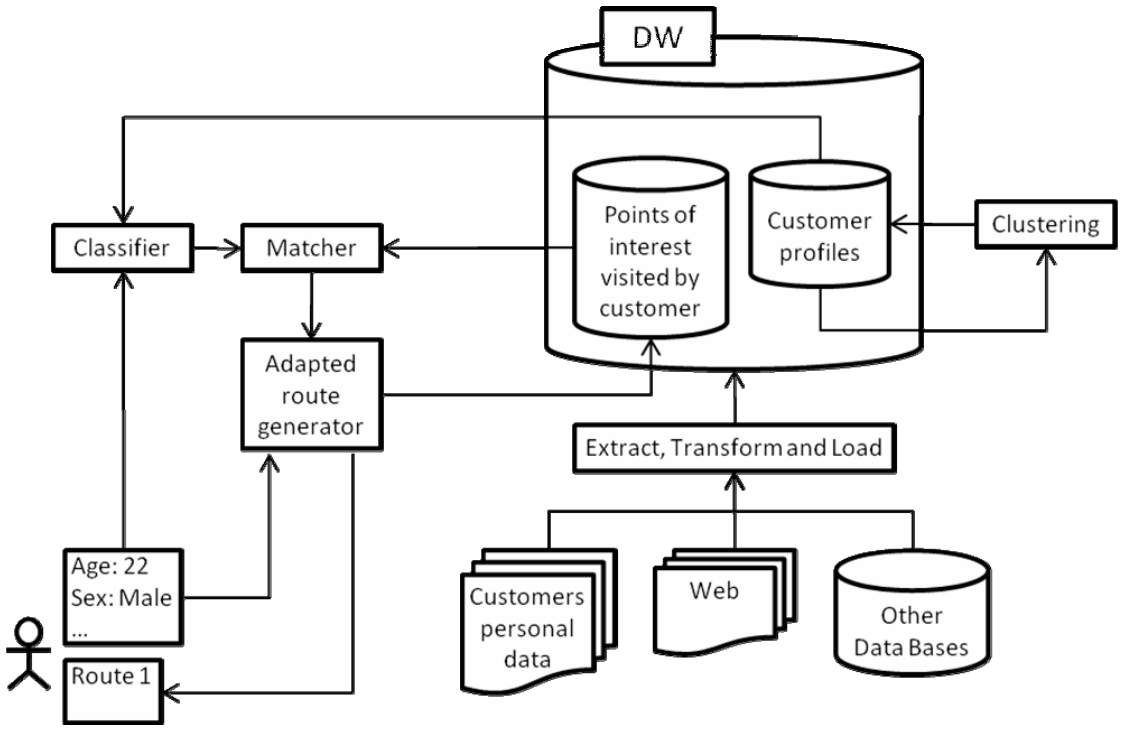

Fig. 4. Data Warehouse oriented architecture

The clustering process has been carried out using Weka [8], a free but powerful tool for this kind of tasks. The customer profile includes several variables. Although they are advised to, the customers can freely fill or not each variable value.

- Age: it is derived from the user's birthday date. In this way the user can be dynamically classified throughout time. If it is not filled, the mean of the already recorded values is considered.
  Sex: it is converted to a unit interval where 0 stands for male and 1 stands for female. If a value is not given, 0.5 is considered as the value.
- Salary: this is a variable which many customers are not comfortable giving it but it can give us much information about the trip. As well as the age variable, the mean value is used when no value is given.
- Marital status: this variable is treated in a similar way as the "sex" is, i.e. it is converted to a unit time interval value where 0 stands for single/divorced/widow and 1 stands for coupled.
- Number of family members: this is an integer variable indicating the size of the family the customer belongs to. When no value is supplied its value depends on the "marital status" variable: if the customer is not coupled the value of this variable will be set to 0 whereas it will be set to the mean of the recorded values in any other case.
- Study level achieved: this variable indicates the study level the customer has achieved. It might forecast the general interest of the customer for the cultural spots (like museums, art galleries…). The possible values (basic education, higher education, Ph Degree…) are transformed to a numerical scale. When no value is given, the mean value of the recorded value is used.
- Overall physical condition: although this variable is more important when planning the trip, once the spots have been selected, it can also have certain importance when searching for spots. What is more, the lower the value for this variable is the higher is the impact over the final interesting spots. There are certain spots which are not suitable for disabled, for instance. As well as the "study level" variable the values for this one are linguistic terms that are converted to numerical values.

- *Classifier*: Once a customer has been registered he is ready for planning the trip but, firstly, he has to selects the spots he wants to visit. One of the objectives of the system is to suggest to the customer the spots he might be interested in, avoiding the bothersome task of browsing the

entire spot data base. The suggestions are perform taking into account the spots which similar customers usually include in their trips. The customer is classified according to the profiles obtained by the clustering process.

- *Matcher*: Once the system has the profile of the user it has to offer him the spots he might be interested in. As it has been stated before, the resultant spots list is based on the spots the user with the same profile have included in their trips. For this reason, the system has to record all the spots the customers include in their trips. Due to the fact that the system recalculates the profiles from time to time, it has to records the spots by user and not by profile. This means that if the classifier process changes the profile a user belongs to, the spots he has include in his trips will be taken into account only in the new profile the customer has been assigned to. This action may result in a change in the list of preferred spots of both customer profiles. To make the system more dynamic an avoid the uniformity, i.e. the same kind of customers always select the same spots because the system always offers them the same spots, some random spots are included in the results. The randomness in the system can be controlled through a variable that determines the percentage of random spots to show to the customer. This variable usually ranges from 5% to 15%.
- *Route adapter generator*: this module is responsible of creating, once gathered the spots the customer wants to visit, the route for the trip. Due to the fact that one of the main objectives of the system is to adapt the route to customer characteristics, the planning of the route is not as easy as finding the shortest path connecting the spots. The process of finding the appropriate route is shown in the next section.

4 Planning the trip

In the previous section the overall architecture of the system has been sketched. In this one we will focus on the process of generating the route which best fits to the customer.

Once the customer has selected the spots he wants to visit, the next step is to find a path connecting them. Usually, when the user does not know much about the area to visit, he gets a map and tries to find the shortest path between then. If he has not got a map, he can also use some of the online tools (Google Earth/Map, Microsoft Virtual Earth...) which will also give him the shortest path between the spots. The shortest path is not always the best option when visiting a city, for instance. Moreover, in certain cases, like when dealing with disabled, the shortest path is simply not valid.

For this reason, one of the fundamental parts of the system is the process of collecting information about the area to visit. It is a laborious task because it has to be done manually. Apart from the location of the spots, information about how to go from every spot to the rest has to be recorded. We define a matrix containing these direct paths. In fact, there can be more than one direct path from a spot to another one. We are not interested in having the entire street map of the area to visit because we are not interested in calculating the shortest path between the spots. Instead, we have different descriptions, annotated with metadata, about how to reach from a spot to another one.

The users of the system play an active role in this part of the system because they are encouraged to add more direct paths from the set of spots as well as defining new spots. When defining new paths between spots it is necessary to specify their metadata. This information is available to the rest of the users and can be used for generating more personalized routes.

This fact implies that the elements of the matrix have to be more than a simple integer indicating the distance between spots. Actually, every element of the matrix is a vector containing a value for each of the metadata properties:

- Distance: this is a usual variable of this kind of problems which indicates the distance, in meters, between two spots.
- Time: this variable indicates the estimated time, in minutes, which is needed to go from a spot to another one.
- Transport: this variable indicates if the path implies to take a public transport (bus or metro). It is expressed using linguistic terms.
- Mean unevenness, Maximum unevenness: knowing this values we can adapt the route to the physical condition of the customer. The mean unevenness is useful for adapting the route to the physical condition of customer whereas the maximum unevenness will be used for discarding routes for disabled. They are expressed in meters.
- Roughness: this variable indicates the kind of surface the path goes by. It is useful when dealing with disabled. It is expressed using linguistic terms.
- Shopping: this variable indicates the shop density the path goes by. It is expressed using linguistic terms.

- Shade: this variable indicates the amount of shade we can find along path. Depending on the forecast conditions and the season it is possible to automatically adapt the route taking into account this aspect. It is expressed using linguistic terms.

Once the customer has selected the spots he wants to visit, finding the paths which best fits his personal characteristics is not a direct process. Due to the fact that we are dealing with a graph containing multiple arches connecting its nodes and that the arches have several associated values it is not possible to use the traditional commerce traveller problem approaches for solving the problem. Although it might not be the best technique, we have adopted a pseudo random based solution. This solution is time cost effective and allows us to deal with multiple users at the same time without overloading the system.

Basically, the process consists on pseudo randomly generate hundreds of possible routes connecting the spots given by the customer, evaluate them according to the selected paths metadata and the customer characteristics and offer him the routes with highest values. They are not actually pure random routes. To enhance the performance of the system the random route generator processor tries to connect every POI with the nearer ones.

For the evaluation process, the user has to specify the importance every link property has for him. By default, these values are set to the mean values given by the users in the same cluster the customer belongs to. In this way, the system only has to aggregate the values of each direct path in the route (average for the unevenness, shopping and shade factors and sum for the rest) and then multiply the resulting links' metadata parameters vector by the customer preferences vector, obtaining a unique value indicating the fitness of the route to the customer. This value is transformed to a percent scale for making sense for the clients.

5 The Granatum project in action

In order to clarify the process summarized in the previous sections we are going to present a detailed example in this one. This example covers all the steps a client has to perform since he is registered in the system until he obtains a touristic route adapted to his preferences and characteristics. It is based on a Web

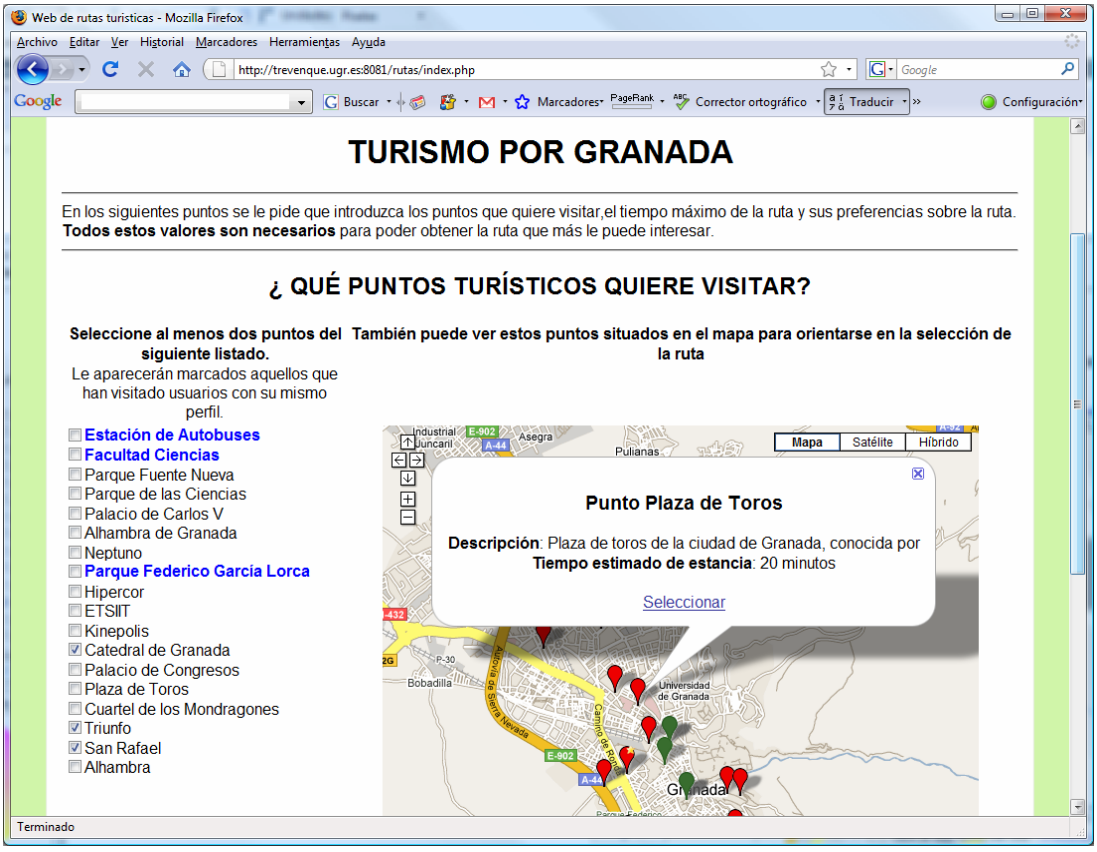

Fig. 5. Web form for selecting the Point Of interest.

site we are developing as a prototype for the Granatum project, granted by the Andalucia Research Program for improving the quality of the touristic services in this region of Spain.

Once the customer has been registered and he has recorded his personal characteristics (see section 3) he has basically three options: see the predefined routes, create a customized route or generate an automatic personalized and adapted route.

5.1 Predefined routes

See the predefined routes the system administrator has defined. These routes are static set of POIs and links between those POIs. The routes presented in this option represent the typical routes any touristic guide offer to the tourist. This is the quickest way the customer can get a general idea of the location he is going to visit.

5.2 Customized routes

If the user is interested in generating a more personalized touristic route he can use this option. It consists on the following steps.

5.2.1 Selecting the Points Of Interest

Once the user has selected this option the first form allows him to select the POIs he wants to visit. As detailed in section 3, one of the fundamental features of the system is to ease the task of finding the most interesting spots for the tourist. In this case, the web form presents the list with all the POIs in the system where the most interesting POIs for the user are emphasized (see figure 5). Also, a map of the region to visit is shown and where the POIs are located by a set of icons. The icons in the map representing the most interesting spots for the user are also emphasized.

Every time a personalized route is produced all the POIs in the route are recorded in form of historical <POI, User, Date> tuples. This way, when the system has to determine the most interesting spots for a client it only has to classify him in a cluster and fetch a list of the most visited POIs by the users who belong to the same cluster. The visited POIs are associated to the users due to the fact that every night the clustering and classification processes are carried out and some users could have been changed to another cluster. Thus, the POIs visited by the users always have influence in the current cluster the users belong to.

5.2.2 Determining the user preferences

After selecting the POIs the user is interested in a second step has to be performed prior of generating the route. In this step the user has to specify his preferences about the route parameters. In figure 6 is illustrated how the user gives his preferences for each parameter through a set of combo boxes. The possible values for each preference are “Indifferent”, “Little important”, “Important”, “Very important”, “Essential”.

Depending on the kind of information the parameter represents it can have a direct or inverse relation with respect to the score of the route. Importance in direct parameters like “shops” or “shade” means that the higher the value the link has for this parameter the higher is the score of the link. On the other hand, the importance in inverse parameters like “time”, “distance”, “public transport”, “Roughness” or “unevenness” implies that the lower the value of the link for this parameters the higher its score. The relation of each parameter is established by the system administrator when defining them.

Fig. 6. Web form for selecting the Point Of interest.

By default the preference value for each parameter is set at the most usual value the users with similar characteristics have used for that parameter. As well as with the process of selecting the most interesting POIs for the user this step is based on the user clustering/classification processes. Thus, the preferences of the users are also stored in the system in form o tuples.

The preferences values for each parameter can be manually modified by the user or automatically by the system itself. According to a set of rules and the information provided by some external data sources. We can say that, for instance, the “shade” parameter is modified according to the rules in table 1 and the data automatically obtained from the Spanish National Weather Institute.

Table 1. Rules for automatically modifying the importance of the "shade" factor according to the weather forecast.

	Spring	Summer	Autumn	Winter
Unknown		+1		-1
Sunny		+2		-2
Cloudy	-1		-1	-2
Windy				+1
Foggy	+1	+1	+1	+1
Rainy				
Snowy	+2	+2	+2	+2

Although it appears to be a complex task for the user he always has the possibility of accept the parameters values proposed by the system and press the continue button.

5.2.3 Generating the personalized routes

Once the system has the POIs and the preferences of the user a list of routes is generated according to the process detailed in section 4 (see figure 7). The personalized routes are sorted according to their score with respect to the preference values given by the user.

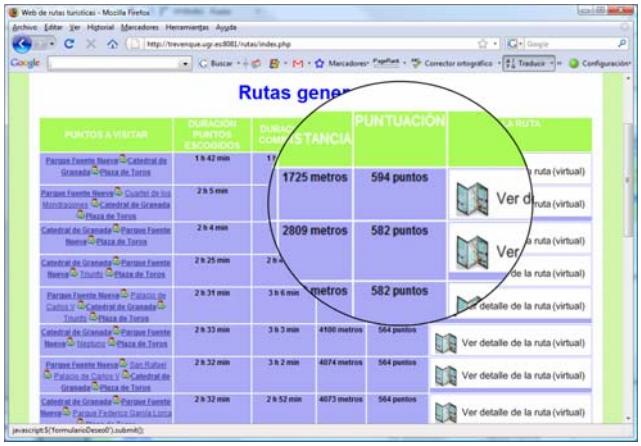

Fig. 7. Personalized routes ranked by the fitness to the tourist preferences.

Due to the fact that all of the POIs in the system are linked with the rest it can occurs that most of routes include POIs which the user has not included in his wish list. This means that the system has generated a route which goes past by some POIs in the system, although the user has not initially shown interest for them. All the intermediate POIs are incorporated to the route in the same way as the POIs in the wish list so their information is completely available for the tourist.

The defined links represent the nodes defined by some user in the system. They have associated a description indicating how to go from one of the node to the other as well as a list of metadata values (distance, estimated duration…).

The virtual links connect every node in the graph with the rest. This kind of links has no more information than the estimated distance between the nodes they connect. Actually, the estimated distance corresponds to the euclidean distance between the nodes multiplied by a factor. This factor represents the extra distance the tourist has to cover due to obstacle between the nodes (buildings…) and it is set by the system administrator. For the sake of simplicity, only the virtual links connecting neighbors nodes are shown in figure 8 (dotted lines).

Let suppose the tourist includes the A and B nodes in his wish list. The system generates two different routes which connect those nodes: route A and B. Depending on the preference values the user gives the former or the latter is given as the route which better fits the tourist preferences. The route A is better than the route B when the tourist prefers to go by areas with shops although it implies to cover more distance. On the other hands, the Route B is better than route A when the tourist prefers the shortest alternative.

The latter case includes in the route a virtual link. In this case, the system has no information on how to reach from node C to node D. The user is informed about this issue but the route can be produced as well as the routes not including virtual links.

5.3 Automatic personalized and adapted routes

In the previous section the process of creating personalized routes and adapting them to the tourist preferences has been explained. The user has also the possibility of generating a set of routes without having to express the POIs wish list and his preferences. In this case a number of POIs – the most included by similar tourist in their routes – are selected by default and the routed adaption to the tourist preferences are carried out taking into account the preferences given by the users with similar characteristics.

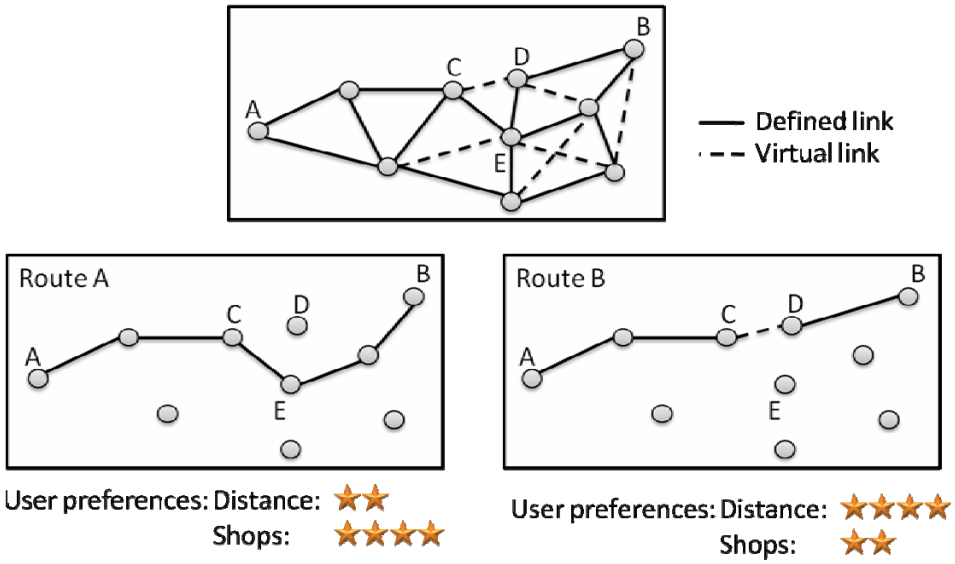

Fig. 8. Route elaboration according to tourist preferences.

By mean of this option the tourist can obtain a list of routes adapted to his characteristics and preferences without having to follow the steps detailed in the previous section. He only has to press one button.

5.4 Viewing the routes

When the user select any of the produced routes he access to the detail of the route. This detail can be obtained in form of text, map or some kinds of mapping application files format.

The text version of the route is an ordered list of the POIs included in the route where all the information about the POIs and the links connecting those POIs is exposed.

The route can also be seen over a map. All the information about the POIs is presented to the user when he selects any of them.

The system is also capable of generating the route in various formats for some mapping applications like Google Earth© or GPS navigation systems like TomTom©.

If desired, the user can save the resulting routes for a later review.

6 Conclusions

In this paper we have presented a DW architecture which helps users on finding the most interesting places for visiting in a trip as well as finding the route which better fits to their characteristics and other environment condition.

We have shown how Data Mining techniques, which are usually employed in business environments, can be successfully used in the tourism area. In fact, the clustering/classification approach coupled with DW functionalities has been presented as a powerful technique for anticipating the users' requests and offering them the information they are looking for.

Acknowledgements

This work has been supported by the Andalucía Research Program under project GR2007/07-2 and by the Spanish Research Program under project TIN2005-09098-C05-03.

References

[1] Adler, J.L., Blue, V.J., 1998. Toward the design of intelligent traveler information systems. Transportation Research Journal. Vol. 6. Pp. 157-172.

[2] Ambite, J.L., Barish, G., Knoblock, C.A., Muslea, M., Oh, J., Minton, S., 2002. Getting from here to there: interactive planning and agent execution for optimizing travel. Eighteenth national conference on Artificial intelligence. Alberta, Canada. Pp. 862-869.

[3] Araque, F., Carrasco, R. A., Salguero, A., Delgado, C., Vila, M. A., 2007b. Fuzzy Integration of a Web data sources for Data Warehousing. Lecture Notes in Computer Science (Vol 4739). Springer-Verlag.

[4] Araque, F., Salguero, A., Delgado, C., 2007. Monitoring web data sources using temporal properties as an external resources of a data warehouse. ICEIS. 28-35.

[5] Atlas virtual, Government of Catalonia, Spain, 2008. http://pandora.icc.cat/cas/

[6] En Vivo, Government of Andalucía, Spain, 2008. http://envivo.andalucia.org/

[7] Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J. Knowledge Discovery in Databases: An Overview. G. Piatetsky-Shapiro, W.J. Frawley eds. Knowledge Discovery in Databases pp. 1-31, The AAAI Press.