Automatic Classification of Location Contexts with Decision Trees

Maribel Yasmina Santos and Adriano Moreira

Department of Information Systems, University of Minho, Campus de Azurém,
4800-058 Guimarães, Portugal
{maribel, adriano}@dsi.uminho.pt

Abstract. Location contexts are geographic regions, with well defined boundaries, that can be used to characterize the context of the persons lying inside them. In this paper we describe a process that exploits the increasing availability of geographic data to automatically create and classify location contexts. The proposed process generates new geographic regions from a database of Points Of Interest through the use of spatial clustering techniques, and classifies them automatically using a decision tree based method. Some preliminary results demonstrate the validity of this approach, while suggesting that a richer geographic database could produce location contexts of higher quality.

1 Introduction

The wide availability of wireless local area networks and mobile cellular networks, as well as the improved capabilities of the current mobile devices, are enabling the development of new and advanced context-aware applications. Within this class of applications, the location-aware applications are those that are able to adapt their behaviour accordingly to the location of the user. A key aspect of the location and context-aware applications is the easy access to the context of their users. One simple way of describing the context of a user is to describe the place where she is at in a particular instant of time. The most usual and elementary form of context is the current position of the user, described as a pair of geographic coordinates, or the location, described as an address or other symbolic reference. However, the description of the context can be enhanced if it includes other elements about the surroundings of the user and/or a richer description of the place, such as its type and current state (places change with the time).

One possible approach to characterize the context of a user, in particular in what concerns its location, is to gather data in real time from a set of sensors (e.g. a Global Positioning System - GPS - receiver), and let the context-aware application itself infer the context of the user [1, 2]. Another approach, which can be combined with the previous one, is to obtain the position of the user, in a geographic referential, and then use it to query a geographic space model where the geographic space is modelled as a set of pre-characterized regions – the location contexts [3]. In this approach, the description of the context of a particular user can be enhanced with the description of the

PREPRINT
Proceedings of the Conference on Mobile and Ubiquitous Systems, Guimarães, Portugal, 29-30 June, 2006

regions (location-contexts) that include the current position of the user. Examples of location contexts are city centres, leisure areas, shopping areas, rural areas, etc.

The major disadvantage of the Space Models approach is that space models must be created for the entire geographic scope of the location-aware applications. Although this can be easily done, manually, for small regions, such as a single town, it is almost impossible to do that for larger areas due to the required manual effort [4]. The work described in this paper addresses the problem of creating space models for large areas by proposing a method to create geographic location-contexts automatically.

The basic idea exploited in this work is that the increasing availability of geographic data enables the automatic identification of regions with particular characteristics, which can be automatically classified as location-contexts of a certain type.

In the following section we describe how commercially available geographic databases can be used as a source of data for the automatic creation of a space model based on location-contexts, and describe the proposed process to do it. In section 3, we describe how a density based clustering technique combined with an algorithm to calculate the concave hull of a set of points can be used to generate new regions from a database of Points Of Interest. These newly created geographic regions are then automatically classified, as described in section 4, to produce a set of new location contexts. Finally, we present our conclusions and identify trends for future work.

2 Extracting location contexts from geographic databases

Commercially available geographic databases usually contain a set of geographic features about a certain region. These often include the administrative boundaries of countries, districts and municipalities, the geometric representation of rivers and roads, the position of cities, airports, railway stations and other point features usually known as Points Of Interest (POIs). Besides its initial purpose, all that information can also be used to characterize a place and, therefore, has the potential to feed a process that automatically creates and classifies geographic regions. In these databases, urban areas are characterized for having a high number of features (streets, addresses, postal codes, hotels, restaurants, etc.) per unit area, whereas in rural areas the spatial density of such features is typically much lower. In this and the next sections we describe how a set of geographic data can be used to create a space model based on location contexts. Our starting point will be a set of geo-referenced POIs obtained from a commercial geographic database.

The proposed process includes three major stages, and is illustrated in Fig. 1. The first two stages are oriented towards the identification of geographic regions that emerge from the available data, and involve the grouping of the POIs (first stage) and the calculation of the geometric boundaries of these groups of points (second stage). In the third stage, the identified regions are classified automatically.

The first stage, based on the identification of groups of points, exploits the fact that POIs are geographically distributed in clusters. In urban or highly populated regions, the spatial density of the POIs is high, while in sparsely populated areas the density of points is much lower. To identify these different regions, a spatial density-based clus-

tering technique was adopted. The detailed process and the achieved results are described in section 3.

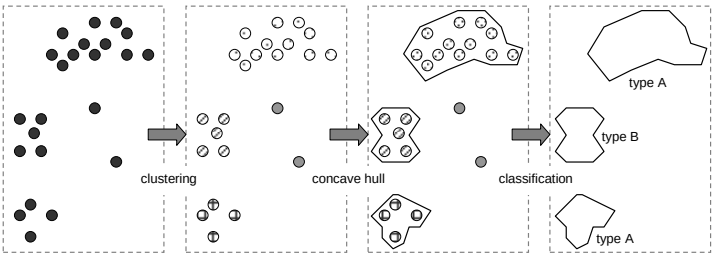

Fig. 1. Generating location contexts from POIs

Once the groups of points are identified, the calculation of the boundaries of the areas containing each group of points defines the new regions. The approach used to calculate these boundaries is also described in section 3.
The third stage, where the regions are classified, is described in section 4, and uses a decision tree data mining algorithm.

2.1 The Points-Of-Interest Database

All the results presented in this paper were obtained from a database of POIs that was acquired from TeleAtlas, a major geographic data supplier. This database contains 18 914 POIs for the Portuguese continental territory, distributed between 55 different categories (airport, bank, pharmacy, etc.). From the original database, a sample of the POIs for a limited region of Portugal was extracted, with a total of 2 549 records. This was the base for all the following steps. The relevant data in each record includes a unique identifier of the POI, the geographic coordinates of the POI in the form of a latitude-longitude pair in the WGS84 datum, and the category of the point.

2.2 Data Mining and Decision Trees

Knowledge Discovery in Databases is associated with the exploration of large amounts of data with the aim of finding trends and patterns in data. This process integrates five steps: data selection, data treatment, data pre-processing, data mining and interpretation of results, which aim to support the decision-making processes [5].

In the Data Mining step, different tasks can be performed and several techniques can be applied in the execution of a specific Data Mining task. Some of these tasks include *classification*, *clustering*, *association*, *prediction* and *estimation*. The performance of each technique depends on the task to be carried out, the quality of the available data and the objective of the discovery. The most popular Data Mining algorithms include *neural networks*, *decision trees*, *association rules* and *genetic algorithms*. In this work *decision trees* were used to perform a *classification* task [6].

Decision trees are obtained looking at particularities of the analysed data, trying to split it in different classes that constitute the output of the learning process. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent the different classes in which the several records can be classified.

A given tree can be transformed into a set of rules of the form `If X Then Y`, where `X` represents a subset of the attributes presented in the analysed data set, and `Y` represents one attribute of the database not present in `X`. For each path, from the root to a leaf node, one rule is obtained. Each attribute-value pair along with a given path forms a conjunction in the rule antecedent ("`IF`" part). The leaf node holds the class prediction, forming the rule consequent ("`THEN`" part). One of the advantages of the use of decision trees in Data Mining tasks is that the knowledge acquired by them can be easily understood and used by humans, allowing their use not only in automatic classification tasks but also in the comprehension of how the decisions are taken.

3 Identification of Location Contexts

In this section we describe how the new geographic regions are created from the geographic database. This involves the two stages described in the previous section: grouping the points and the computation of the boundaries of the new regions.

3.1 The Clusters Identification

The identification of the regions was first performed by clustering the points that, due to their geographic proximity, formed a kind of "cloud" in the geographic space. For this task, a spatial density-based clustering algorithm was used – the Shared Nearest Neighbour (SNN) algorithm [7]. This algorithm is adequate for this particular task due to its capabilities to identify clusters with convex and non-convex shapes, with different sizes and densities, and also due to its ability to deal with noise points. Since no implementations of this algorithm were found, we developed our own code, which is now available for download from the LOCAL project web site (http://get.dsi.uminho.pt/local).

The clustering process was repeated using several values for the *k* parameter of the SNN algorithm, and the obtained results were compared to assess the quality of the created clusters. This comparison was made by displaying the generated clusters on top of satellite images from the Google Earth application [8], and by visually observing the relation between the clusters and the other geographic elements represented in the images. The final value of $k=8$ was chosen as the one that produces the clusters that best represent the corresponding area. Although $k=8$ was found to produce the best clusters for this data set, it is possible that slightly different values have to be used for other data sets, namely for data sets with much higher densities of points. Therefore, further validation on this issue is required, but it depends on the availability of other geographic databases. Fig. 2 shows an example of a set of clusters identified by SNN with $k=8$.

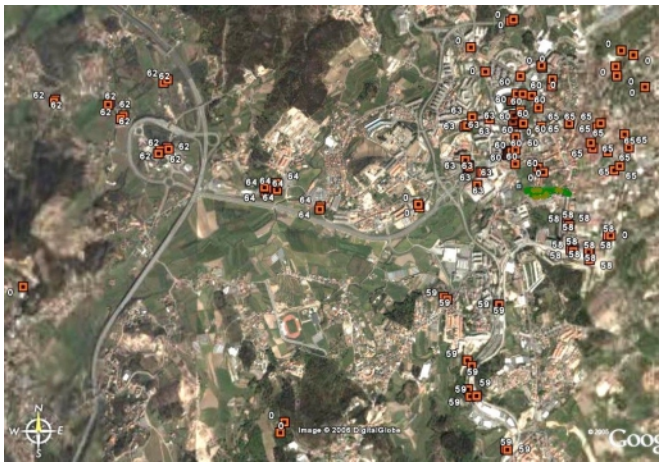

Fig. 2. Some clusters identified by SNN for k=8

3.2 Topology Definition

Once the groups of points were identified, the creation of new regions was achieved by calculating the geometric boundaries of each group of points. For groups of points that occupy a convex shaped area, a convex hull [9] algorithm can be used to calculate the polygon that represents the geometric boundary of that area. However, many of the clusters identified by the SNN algorithm are not convex shaped and, therefore, a convex polygon does not properly represent the area defined by those clusters of points. In these cases, an algorithm that could create both convex and non-convex polygons from a set of points would be a better solution. Unfortunately, a literature review in this field did not produced evidence that one such algorithm exists. The solution was to develop a new algorithm that generates both convex and non-convex polygon from a set of points (concave hull algorithm), and another algorithm to add a "buffer" around that polygon. The description of the concave hull algorithm is out of the scope of this paper and can be found in [10]. Fig. 3 shows an example of the polygon computed by the concave hull algorithm for one particular cluster of points, and the "buffer" (another polygon) that better represents the area occupied by the given points.

4 Automatic Classification of Location Contexts

In this section we describe how to build a model to automatically classify the clusters generated by the SNN algorithm (third stage of the process). For this task, an approach based on Decision Trees was used. First, the clusters were manually classified by displaying them on top of satellite images, as shown in Fig. 2. Looking at the satellite images and looking at the characteristics of each region, a class was manually

assigned to each region. This manually classified data was then used to train a decision tree in order to create a model that allows the automatic classification of the regions. The set of 24 classes used to manually classify the clusters is shown in Table 1 (first column), and were created taking into account the available geographic data.

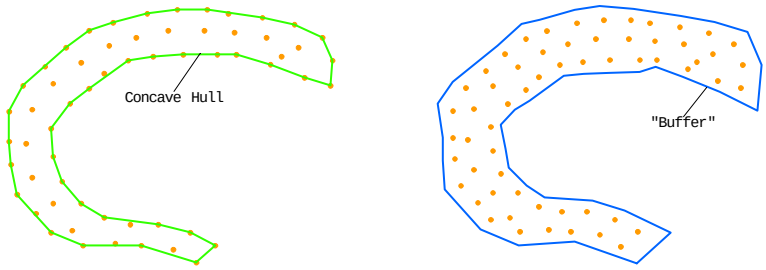

Fig. 3. The computed boundary of a set of points (concave hull and buffer algorithms)

4.1 The classification process

The classification process adopted in this work includes the five steps traditionally associated with the knowledge discovery process, namely: i) data selection, ii) data treatment, iii) data pre-processing, iv) data mining and v) interpretation of the results.

In the data selection phase, the regions previously created were analysed in order to identify attributes that are not relevant to the data mining step. The data treatment phase is concerned with the identification of noise or missing data fields, in order to identify any inconsistencies present in the data.

The data pre-processing phase has as objective the reduction of the data set, in terms of columns or records available for analysis to the data mining algorithms. Another objective of this step is the transformation of the type of the attributes, in order to be possible their analysis by specific algorithms. As this work addresses the use of the C5.0 algorithm (version available in [11]), in the induction of a decision tree, and as this algorithm can handle numeric and alphanumeric values, no transformation was required. Also, at this step, the data available for analysis was divided in two data sets, the training and the test data sets. The first one was used in the identification of the model, and the second one in its validation. In the data mining step the C5.0 algorithm was used to identify a model that allows the automatic classification of the location contexts. This model was validated in the interpretation of the results phase, in order to verify its accuracy in the classification of unknown data.

4.2 Training and Testing of a Decision Tree

The process of analysis was initiated by the exploration of the data set available, in order to identify the relevant attributes, and also to identify noise or other problems present in the data.

The data mining application used in this work was the Clementine Data Mining System v8.0 [11]. Clementine is based on visual programming. For the execution of a specific task or set of tasks on data, the construction of a stream is required, in which each operation on data is represented by a node.

The first generated table is presented in Fig. 4 and shows an extract of the several attributes available in the data set. Each record represents a cluster, and each cluster is identified by: i) a `ClusterID`, ii) the number of POIs present in the cluster (`points`), iii) the geographic area of the cluster (`clusterArea`), iv) the manual classification given to the cluster (`clusterClass`), and v) the number of points of a specific type present in the cluster (for example, cluster `B1` contains `1` point of the type `Car Repair Facility`, `0` points of the type `Petrol Station` and so on). The analysed data set includes 55 different types of points (POIs).

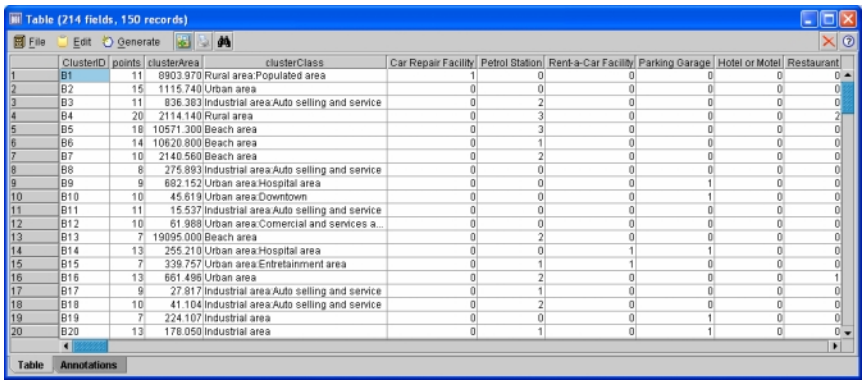
Table (214 fields, 150 records)

File Edit Generate

	ClusterID	points	clusterArea	clusterClass	Car Repair Facility	Petrol Station	Rent-a-Car Facility	Parking Garage	Hotel or Motel	Restaurant
1	B1	11	8903.970	Rural area:Populated area	1	0	0	0	0	0
2	B2	15	1115.740	Urban area	0	0	0	0	0	0
3	B3	11	836.383	Industrial area:Auto selling and service	0	2	0	0	0	0
4	B4	20	2114.140	Rural area	0	3	0	0	0	2
5	B5	18	10571.300	Beach area	0	3	0	0	0	0
6	B6	14	10620.800	Beach area	0	1	0	0	0	0
7	B7	10	2140.560	Beach area	0	2	0	0	0	0
8	B8	8	275.893	Industrial area:Auto selling and service	0	0	0	0	0	0
9	B9	9	682.152	Urban area:Hospital area	0	0	0	1	0	0
10	B10	10	45.619	Urban area:Downtown	0	0	0	1	0	0
11	B11	11	15.537	Industrial area:Auto selling and service	0	0	0	0	0	0
12	B12	10	61.988	Urban area:Comercial and services a...	0	0	0	0	0	0
13	B13	7	19095.000	Beach area	0	2	0	0	0	0
14	B14	13	255.210	Urban area:Hospital area	0	0	1	1	0	0
15	B15	7	339.757	Urban area:Entretainment area	0	1	1	0	0	0
16	B16	13	661.496	Urban area	0	2	0	0	0	1
17	B17	9	27.817	Industrial area:Auto selling and service	0	1	0	0	0	0
18	B18	10	41.104	Industrial area:Auto selling and service	0	2	0	0	0	0
19	B19	7	224.107	Industrial area	0	0	0	1	0	0
20	B20	13	178.050	Industrial area	0	1	0	1	0	0

Table Annotations

Fig. 4. Data available for analysis

The distribution present in the data set in terms of the different classes of clusters is represented in Table 1 where it is shown that for the 150 available records, 4 cluster classes are dominant, in terms of the number of times that they appear in the data set: `Urban area`, `Urban area: Commercial and services area`, `Industrial area: Auto selling and service` and `Beach area`. The `clusterClass` with the description '`na`' was assigned to two clusters that, due to their geographic extension, could not be associated to any of the pre-defined classes, and was excluded from the analysis in the data treatment phase.

After the exploration of the available data, it is now possible the selection of the attributes considered relevant to the following steps. The only attribute considered without relevance for the classification process is the `ClusterID`, reason why at this point it was removed from the data set. In the data pre-processing step, and since the reduction of the sample available for analysis is not possible, neither in terms of attributes or records, the data set obtained in the data treatment phase was used to create the `Train_Data` set and the `Test_Data` set, using the proportion of 70% of records for training, and the other 30% for testing. Ideally, the number of records for validation should be greater than the number of records available for training. However, as the number of records available for analysis in this data set is small and as the number of

records available for each clusterClass in not homogeneous, a higher percentage of records for training is required in order to obtain a representative training set.

Table 1. Distribution of the clusterClass attribute

clusterClass	frequency	count
Beach area	10,0	15
Exhibition area	1,33	2
Highway service area	1,33	2
Industrial area	2,67	4
Industrial area: Auto selling and service	13,33	20
Rural area	1,33	2
Rural area: Populated area	0,67	1
Sports area	1,33	2
Suburban area	4,67	7
Suburban area: Auto selling and service	0,67	1
Suburban area: Commercial and services	3,33	5
Suburban area: Entertainment area	1,33	2
Suburban area: Residential area	2,0	3
Suburban area: Train station area	0,67	1
Urban area	14,67	22
Urban area: Auto selling and service	0,67	1
Urban area: Commercial and services	14.67	22
Urban area: Downtown	5,33	8
Urban area: Entertainment area	2,0	3
Urban area: Historic area	1,33	2
Urban area: Hospital area	4,0	6
Urban area: Residential area	4,67	7
Urban area: Train station area	2,0	3
Urban area: Turistic area	4,0	6
na	2,0	3

In the data mining step, the data available in the Train_Data set was used for the training of a decision tree with the C5.0 algorithm. Fig. 5 shows a partial view of the obtained model.

In this model, the attributes that were considered relevant by the C5.0 algorithm were: the number of points of the cluster (points) and several types of POIs, namely, Beach, Hospital/Policlinic, Important Tourist Attraction, Rest Area, Restaurant, Courthouse, Car Dealer, Exhibition Centre, Railway Station, Car Repair Facility, Hotel or Motel, Cash Dispenser, Pharmacy, Government Office and Museum.

The Test_Data set was then used to verify the performance of the obtained model in the classification of unknown clusters. With the first model, a confidence of 37% was obtained, which is a low value for a classification model. This result calls for an additional analysis of the data.

The work undertaken until now allowed the identification of two major problems present in the data. One is associated with the reduced number of records available in the sample manually classified and the other is related with the fact that the sample is not homogeneous in terms of the representation of each clusterClass, as shown in Table 1. One of the tasks that can now be carried out is the balancing of the data set. Two balancing strategies are possible: by reduction of the number of records (which is not possible in this data set due to the reduced number of records) or by "boosting" of the number of records, which means replicating the records available for the classes with less elements (this task is done automatically by Clementine).

Using the boosting option for gathering a well-balanced data set, the results obtained were very good. The performance of the identified decision tree increased enormously, reaching the value of 92% in the `Test_Data` set.

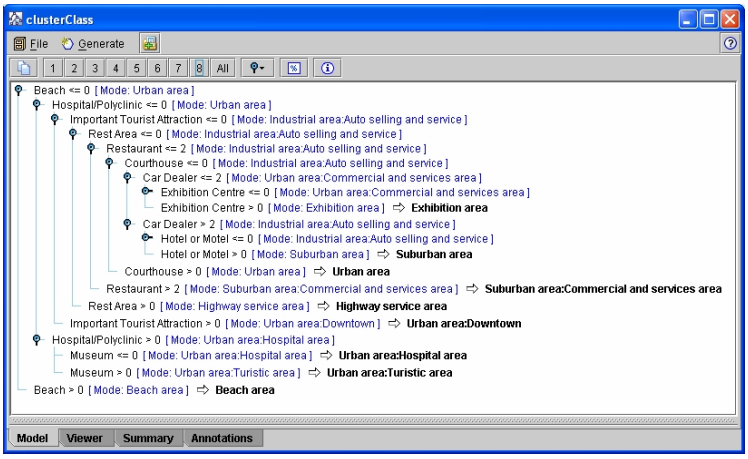

Fig. 5. Partial view of the obtained model

Fig. 6 shows the analysis node with the results and a subset of the rules obtained by the decision tree. In terms of streams, Fig. 6 also allows the verification of the three streams that were needed in the analysis of the data set with the boosting option: i) the first one (the stream on the top of the figure) was used in the data selection, data treatment and data pre-processing steps; ii) the second one (at the middle of the figure) was used in the data mining step and iii) the third one (at the bottom of the figure) was used in the interpretation of the results step.

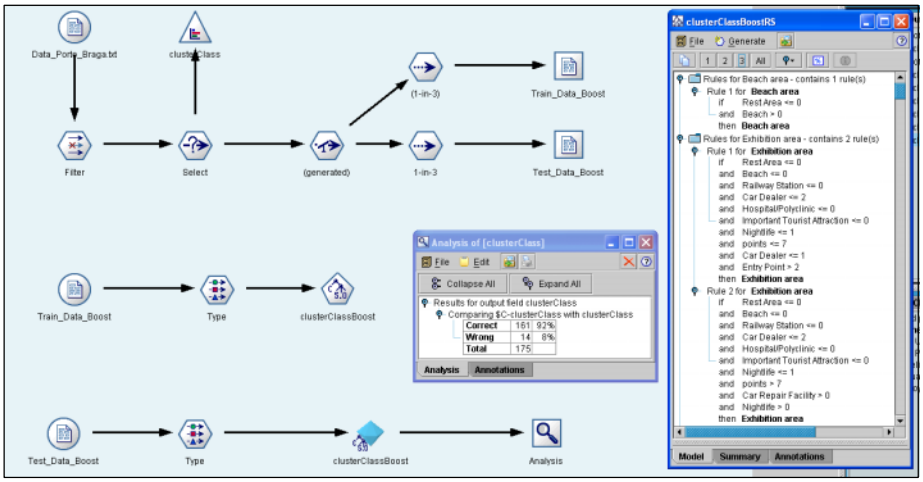

Fig. 6. Identification of a classification model with a boosted data set

5 Conclusion and Future Work

This paper presented the analysis of a database of POIs with the aim of automatic creation and classification of location contexts. To the best knowledge of the authors, no other work has proposed a similar process. This work started with the clustering of the several POIs using the SNN algorithm. After the clustering process, the polygon of each cluster was defined recurring to the concave hull algorithm.

The classification process started with the manual classification of a set of clusters that were used in the training and testing of a decision tree. Due to the reduced number of clusters manually classified and the heterogeneity of the sample, the confidence of the first model was very low. In order to improve the classification capacity of the tree, the balancing of the sample was done recurring to a boosting technique. After this, a model with 92% of confidence for automatic classification of location contexts was identified. The obtained model can now be used to create a database of location contexts from the geographic database that covers the entire continental Portuguese territory.

Although good results were obtained in terms of precision of the model, future work includes the increasing of the initial data set, adding more records manually classified and also a more balanced sample.

References

1. S. Pradhan, C. Brignone, J.-H. Cui, A. McReynolds, and M. T. Smith, "Websigns: Hyperlinking physical locations to the web," IEEE Computer, Vol. 34, No. 8, pp. 42–48, Aug. 2001
2. Anthony La Marca, et al, "Place Lab: Device Positioning Using Radio Beacons in the Wild", Proceedings of the Pervasive 2005, Munich, Germany, May 2005
3. Barry Brumitt, Steven Shafer, "Topological World Modeling Using Semantic Spaces", UBICOMP 2001 Workshop on Location Modeling for Ubiquitous Computing, October 2001
4. R. K. Harle, A. Hopper, "Dynamic World Models from Ray-tracing", Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications (PerCom'04), Washington, DC, USA, 2004
5. U. M. Fayyad, et al., eds. Advances in Knowledge Discovery and Data Mining, The MIT Press: Massachusetts (1996).
6. J. Han, Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2001)
7. Levent Ertoz, Michael Steinbach, Vipin Kumar, "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data", Proceedings of the Second SIAM International Conference on Data Mining, San Francisco, CA, USA, May 2003
8. Google Earth application: http://earth.google.com
9. F. P. Preparata, S. J. Hong, "Convex hulls of finite sets of points in two and three dimensions", Communications of the ACM, Vol. 20, No. 2, pp.87-93, February 1977
10. Adriano Moreira, Maribel Yasmina Santos, "A k-nearest neighbours approach for the calculation of the concave hull of a set of points", Submitted to the International Symposium on Advances in Geographic Information Systems – ACM-GIS'06, USA, 2006
11. SPSS, Clementine, User Guide v8.0, SPSS Inc. (2004)