ChoreMAP: Extracting And Displaying Visual Database Summaries Tool

I. Cherni
LTSIRS, Institut des Sciences de Gestion of Tunis, Tunisie, LIRIS,Institut National des Sciences Appliquées de Lyon, France
email: c.ibtissem@hotmail.fr

S. Faiz
ISAMM, de la Manouba, Tunisie
LTSIRS, ENIT de Tunis
email: sami.faiz@insat.rnu.tn

R. Laurini
LIRIS, INSA of LYON, Villeurbanne, Lyon, France
email: robert.laurini@insa-lyon.fr

M. Warghi
Faculté des Sciences Juridiques, Economiques et de Gestion de l'Université de Jendouba, Tunisie
email:mariem2012.warghi@gmail.com

Abstract- **Traditional cartography is an essential tool to describe the facts and the relations concerning a territory. Expert users are usually satisfied with the expressive power of traditional mapping, when it deals with simple cases. But in some complex cases including a large number of data, the expert users need a map which stresses the most important aspects rather than have several maps with a high level of details. It is in this context that our search has been launched in order to automatically discover spatial patterns and view based on a spatial database and chorems. This paper presents the project focusing on the extraction subsystem of salient patterns that will be encoded with an extensible markup (XML)-based language called chorem markup language (ChorML) and then be viewed as visual summaries by visualization subsystem.**

Keywords- Chorem, geographic databases, summaries, geographic knowledge, data mining.

I. INTRODUCTION

The geographic databases contain the necessary information to the understanding of our environment; thus they help us make decisions related to our environment. Of course this information needs a clear and easy representation to be understood. Therefore, to be able to make decisions we need maps that give a synthetic vision of a whole which integrates visual summaries that easily describe and explain the important information.

Thanks to chorems, we are able to represent the knowledge we have about a certain place in a very simple and clear way. It is also thanks to their abilities to symbolize and encapsulate a methodology and its corresponding interpretation. We can also show climatic, geographic, economic, geologic**,** and agronomic situations, based on their spatial and statistical temporal context, by combining many chorems to form a chromatic card.

In this paper, we describe the research carried out within an international project, launched among several research institutions. The project is meant to define cartographic solutions able to better represent geographic information extracted from (spatial) database contents, which refer to both static objects and evolutionary phenomena.

It is by studying chorems that we were motivated to contribute to the development of a system to represent situations, like those mentioned previously, starting by the results of data mining on a geographic database. The most common situations are those represented in the study of the structure and the dynamics of population, urban concentrations or the interaction between natural and social systems. In these models, there is a problem to define a method to generate chorems that algorithms used to accede a database then extract patterns, which are visualized on a chorematic card.

The objective of this paper is to present the challenges related to the automatic extraction of the most significant elements, and then the creation of the chorems from the geographic databases.

The paper is organized as follows: in the next Section, we propose the model that we study in this paper, and some definitions of chorems are given. Section 3 presents clearly the problem. In Section 4, the architecture of our system is presented. Section 5 presents the extraction sub-system. In Section 6, the computational simulation is made to illustrate the efficiency of the algorithm. Finally, we conclude the paper with a case study in Section 7.

II. RELATED WORKS

An immediate synthesis of data of interest, disregarding details, is a real support for human activity to model and analyze the reality of interest. Such a synthesis may then represent the starting point for further processing tasks aimed at deriving spatial analysis data and represent them in a clear map so that expert users could transfer the principal idea of the problem in a simple way or could satisfy a decision maker. In the remainder of this Section, we describe two cartographic representations: cartograms and chorems. According to the definitions in [7], a cartogram is "a small diagram, on the face of a map, showing quantitative information" or "an abstracted and simplified map the base of which is not true to scale".

While cartograms show values of a single variable at a time, chorems allow designers to assemble into a single map more than one thematic layer, thus, representing the relative importance of a set of objects and phenomena related to each other. According to the definition of the French geographer Roger Brunet, who invented chorems [1], a chorem is a schematized territory representation, which eliminates any detail not necessary to the map comprehension. In this definition, the term schematized means that the most important characteristic is a sort of synthetic global vision emphasizing salient aspects. Moreover, Brunet defines the chorems as "elementary structures of the space represented by graphic model used so that all the spatial configuration rise from the combination, eventually very complex, of simple mechanisms" [2].

Now, a second definition of chorem can be proposed: "a chorem can be seen as a visual way to represent geographic knowledge, and so it can be a tool to summarize geographic databases" [5].

Nowadays, there are many systems to create cartograms automatically like QGIS cartogram creator, Scape Toad, Arc Toolbox, etc., But, there is just one system that is created to generate chorems [11]. ChEViS (Chorem Extraction and Visualization System) is an international project lunched in 2009. It is composed of two subsystems: chorem Extraction subsystem and Chorem Visualization subsystem. This system proposes a new method to construct chorems automatically, especially chorems for each user.

III. PROBLEM STATEMENT

Expert users are usually satisfied with the expressive power of traditional mapping when it deals with simple cases. But in some complex cases including a large number of data, the expert users need a map which stresses the most important aspects rather than having several maps with a high level of details.

So, our objective is to define geo-visualization solutions which can adequately represent the information extracted from geographic data. Visual models based on chorems can interpret and represent spaces, their geographic distributions and their dynamics. The same space can be represented in different ways, but all the corresponding maps will tell the same thing. We cannot change the message, its position, its hierarchy, its network, and all those items expressed in the chorematic map. So in this paper, we answer many questions, such as:

- How can we build chorems from geographical databases?
- Which algorithm is used to elaborate clustering to generate chorems?
- How can we extract knowledge to represent the spatial interaction between clusters?

IV. SOLUTION

This paper is issued from several experimentations concerning chorems, "visual schematic representation of territories". The project is an international one. It was developed jointly by several research laboratories. It aims to define mapping solutions capable of synthesizing the content of geographic databases and represent them in a plain, legible and intuitive way.

Figure 1 shows the architecture ChoreMAP (Chorem Mapping) [14], which is composed of three sub-systems: two extraction subsystems and a display subsystem. The previous figure shows the architecture of the proposed system, which consists of three main levels at both extraction and visualization subsystems.

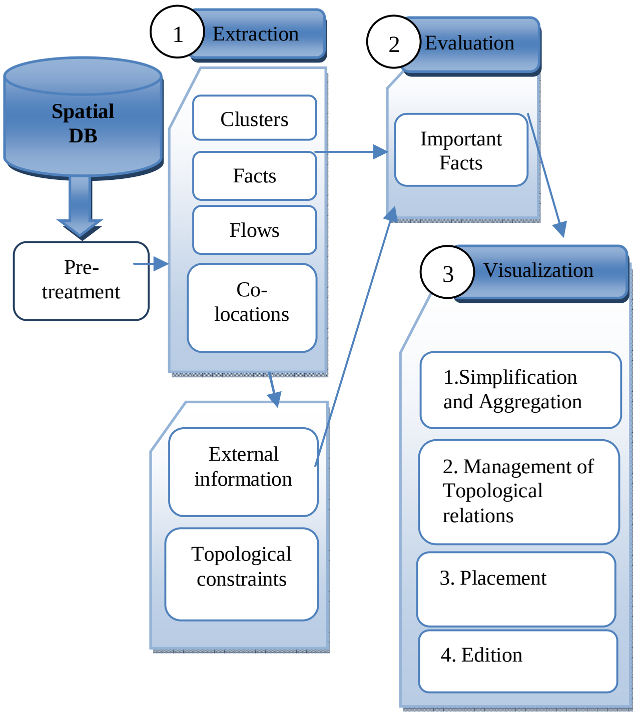


Figure 1. General architecture of the ChoreMAP prototype.

The extraction patterns subsystem is invented to obtain and manipulate information from available data. The extraction of important chorem subsystem manipulates the data generated by the first subsystem. It was strengthened by the integration of space technology of data mining, filtering based on SQL; as to the procedure to reduce the number of patterns, it requires the intervention of an expert taking into account what he wants to show on his card. Thus, the visualization subsystem manages this information by assigning a visual representation in terms of chorems and chorematic cards. It is important to note that the chorem structure is managed by the three subsystems. However, they use and/ or modify the various subparts of chorem structure. In particular, the two extraction subsystems manage chorems conceptual properties, such as name, type, size and coordinated by expressing them by alphanumeric attributes. The sub-display system develops and modifies the geometric shapes of chorems from conceptual properties generated by the previous phase in order to assign an appropriate visual representation. Communication between

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-386-5

levels is based on the multi-level ChorML language. In particular, a suitable level of ChorML is used to correspond to different specifications and to adapt to different useful formats in the running process.

V. SPATIAL PATTERN DISCOVERY

As indicated, the architecture is decomposed into three parts**:** chorem extraction, chorem important extraction and chorem visualization (Figure 1).The application of data mining and spatial data mining algorithms will result in patterns that will be later associated with chorems.

The Chorem Extraction System first transforms the actual data base of available geographic data in order to facilitate the extraction of significant information by spatial data mining. Then, Chorem Visualization System manages this information by assigning them a visual representation in terms of chorems and chorematic maps [5]. Once the patterns are extracted, these results are encoded in the ChorML language by a special subsystem for generating ChorML documents [3], and then visualized.

As previously stated, since there are four types of patterns resulting of the data mining, which seem to be the most interesting in the discovery of chorems, our extraction subsystem is composed of four parts:

- Subsystem of extraction facts.
- Subsystem of extraction flows.
- Subsystem of extraction of clusters.
- Subsystem of extraction co-localization patterns.

In order to encode this knowledge, a special language named ChorML was designed [3]. This language is composed of three important instances**:** First, ChorML 0 shows the database where data mining techniques are going to be applied. At this point spatial data is encoded in GML. Second, ChorML 1 is defined to hold the list of extracted patterns resulted from data mining processes. It contains some additional information such as the elements in the vicinity of the territory, a description of the boundary run on a list of topological and structural relationships between patterns, and finally, at ChorML 2 Geometric shapes have passed through a generalization process and coordinates are translated into a layout format – SVG [4].

A. Facts

A fact is considered the result of one or more queries against the database [9]. A set of rules is defined in order to obtain basic information from the database [4]. To achieve this sub-system, two methods are required: one to analyze data through a user request and another to encode and store the results in a file ChorML.

B. Clusters

Clustering is the method used to group data into classes. Consequently, an object in a cluster has certain similarities with other objects in the same cluster. For example, we could group parcels in a city or by their land use type or group regions by their weather similarities [9]. Clusters are strong candidates to generate chorems [11]. There are many different algorithms for clustering (for example the distribution of clusters). But which one do we use to realize these spots (or subsystem)?.

After a detailed study of methods of data mining, we set the most appropriate algorithm k-means clustering to group the cities that are geographically close and share common characteristics of a set of groups fixed in advance. The proposed method is composed of three modules. It is represented in Figure 2.

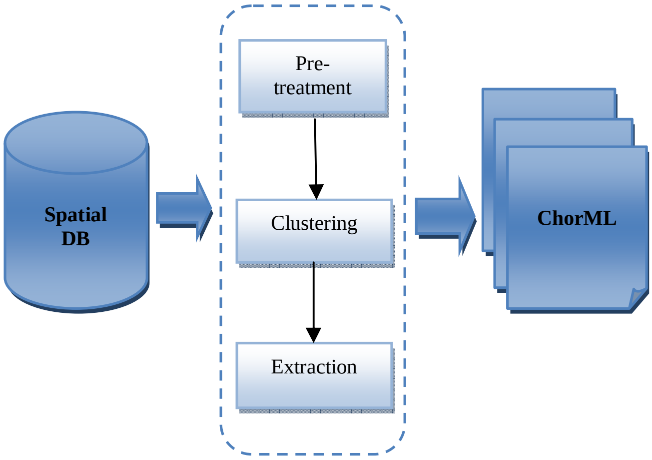


Figure 2. Architecture of sub-system of the clustering extraction

Pre-treatment module: The geographic data are different to those traditional ones. They are stored in the base in thematic layers. In our work we only consider digital data like longitude and latitude to implement the grouping. This module is equipped to create a file with extension “Arff” from data stored in the database.

Clustering module: In this module, we are interested in gathering cities into a set of clusters. In our work, by selecting areas that have common characteristics, we find that the spatial clustering and more specifically the Kmeans algorithm is the most appropriate method. This method allows to extract knowledge that match the pattern structure in the cluster ChorML language with Euclidean distance, Kmeans allows us to select areas with a maximum set of connected points that share the same characteristics. We use weka to affect extraction using Kmeans. Our algorithm is developed:

Step1: Enter the number of clusters (k).
Step2: Choose arbitrary k cities. Affect these cities in the k clusters (cities are the centers of the clusters).
Step 3: Assign each city V cluster Ci Mi center as dist (V, E) is minimal.
Step 4: Recalculate Mi of each cluster (the geometric center).
Step 5: go to step 3 if we just made an assignment.

Copyright (c) IARIA, 2015. ISBN: 978-1-61208-386-5

Extraction module: This module generates a document of clusters from the knowledge extracted by the selection and representation module.

C. Flows

One study showed that three types of flows are the most important [4]: flow path, divergent source flow and well oriented flow. The stream type represents a flow path where the origin and destination are well defined and it may possibly have a geometric shape (for example a large arrow). While the divergent flow type source has a definite origin, the destination is a little uncertain. The destination is a list of different geographical directions. Finally, the convergence has a definite destination, but the origin is a list of convergent geographic directions.

Flows are used to represent the spatial dynamics within a territory. “We consider as flows every movement, material or immaterial, of goods, of people, of information, between different locations”[8]. Flows are generally represented by arrows in the current mapping. We are particularly interested in the flow of goods.

Several authors, like Yann et al. [13] and Tobler [12], find that the flow of goods is mainly related to three factors: (i) the emissivity of the zone i (ii**)**, the attractiveness of zone j and (iii) the inverse function of the distance between the two zones i and j. The distance between the zones has a space factor and an economic separation. Based on these studies, to retrieve related knowledge flows between clusters, we propose a method to study the available quantity of goods. Through the census figures of the city, the consumption and production of products for each city, we can get a good approximation of the movement of goods. This method is done by comparing the production and consumption of agricultural products.

First, to extract the flow of goods, it is necessary to study the quantities available in each city for each product. This method is to subtract the quantities produced: a product with the quantities consumed of the same product. The quantity calculation equation is:

$$Q_t = P_r - P_t * C_i \quad (1)$$

Or

P_t = population in year "t"

C_i = consumption in the year "t" of a product (p) for a person living in the city (V)

P_r = production of a product for the city (V) during the year "t"

Q_t= the quantity demanded (missing) for the product (p) in year t for the city (V).

After the preprocessing module where we store the geometrical shapes in the thematic layers and the descriptive data in a cube data base, we apply the proposed method of goods flow extraction. This method consists of:

- Calculating the quantities of goods available for all cities in the database (using Talend Open Studio).
- Determining the quantities available for each product group from the cluster extraction subsystem.
- Identifying clusters that have seized a quantity above the threshold. These clusters are considered as emitter groups of the product.
- For each receiver group, we compare the amount available for the quantities available for group issuers and the distance between the groups. The selected cluster is the nearest cluster and has an available quantity for the product p.
- Encoding flows in ChorML language.

D. The co-location patterns

Co-location patterns are sets of characteristics of places that are presumed to be close with a certain probability to each other. Co-localization rules seem interesting in creating chorems because they define the organization of objects within the territory with a quantitative accuracy. The results of the extraction modules and those made of clusters are used to determine the relationship between major cities to the user and the resulting groups of the k-means algorithm. For example, if there are commonalities between two geometric shapes, we have a relation of the type 'Touch'. The boundaries of these shapes touch, but their inside does not.

VI. EXPRIMENTS

The inland transport of merchandise is carried by road, rail or inland waterway. According to international definitions, transportation means a flow of goods moved over a given distance and is measured in ton- kilometers.

According to [15] and given the large amount of digital information available in the world, statisticians have the difficult task of ensuring that the trade analysts and others have speed access to accurate business data.

Accurate and up to date production is costly and requires resources that are unfortunately still lacking in many developing countries. That is why the frequency and level of detail of national statistics vary considerably from one country to another. It was often difficult to establish timely and comparable statistics on trade in some of developing countries because these countries do not regularly communicate.

Demographic and economic dynamics of major Tunisian cities is an undeniable reality in the recent reconfiguration of the Tunisian territory. As a basic source of the Tunisian economy, intra-regional trade in food products is considered as the main flow of goods.

Taking into account their importance, and because of the difficulty of their expert analysis, we find it interesting to address this limitation and propose an easier method. In fact, it would be interesting to represent flows on a chorematic map. This provides an easy and synthetic vision as massive data rate in the territory and donated goods will be replaced by forms and symbols easy to understand.

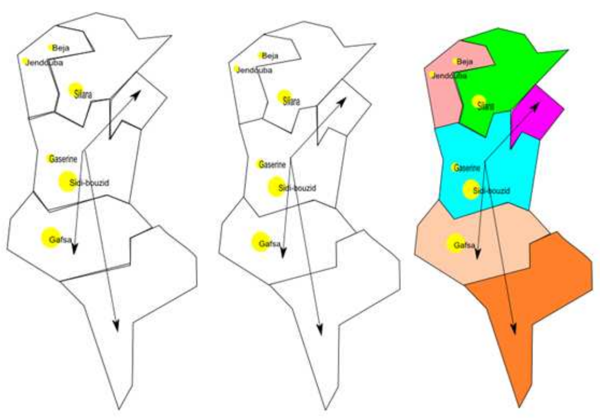
In what follows, we turn to the testing phase where we use as input system a ChorML file from ChoreMAP project [14] that contains the flow of food and agricultural products between regions

A. Generation of chorematic map

In this Section, we present the five main features of our system which are:

- Extraction of the possible groups according to the number entered by the user.
- Extraction of cities issuing a selected product.
- Extraction of the flows of goods between different groups.
- Creating the ChorML document.
- Test results obtained by display on a chorematic map.

We present in the following the results of our experimental study. When we apply the extraction flow method on goods imported from the National Institute of Statistics (INS) in Tunisia and the Regional Offices of Agricultural Development (CRDA) actual data, the result of the experiment is a chorematic map presented in Figure 3, which shows its performance in the presentation of the flow of grain between the main regions in Tunisia.



a. Chorematic card before refinement

b. Chorematic card after refinement

c. Chorematic card after changement of clusters colors

Figure 3. Tunisian Chorem of flows of cereals.

This Chorematic map was generated by using the results obtained from the extraction phase and knowledge encoded ChorML. It describes the quantity transported from one region to another taking into account the proposed threshold.

The menu offers an overview of the most salient freight elements process. It can be used as a decision support. This map shows more precisely chorematic directions grain transportation in Tunisia. It consists of three chorems:

- The territory is divided into six major regions: Northwest, Northeast, Central east, Central west, South east, and Southwest.
- The most productive cereals cities in Tunisia.
- Arrows representing the flow of inter -regional cereals.

With this chorematic card, we can see the existence of the important movement of grain between Tunisian areas. Indeed, the Center west supplies other regions (South east, South west and Center east).

Regions producing cereals are characterized by the relative importance of the rural population and the high proportion of agricultural employment in total employment.

VII. CONCLUSION

Traditional cartography is an essential tool to describe the facts and the relations concerning a territory. Geographic concepts are associated with geographic symbols and graphic symbols help the readers understand immediately the visualized data.

The representation of chorems provides us with the best interpretation of problems. It is in this way that we can obtain all what we need: from the young pupils who want to learn geography, up to the researchers who investigate new forms of communication.

If we want to process other types of information without being confused, we propose the concept of chorem layers to present various phenomena and we also offer the superposition.

In this paper, we present our project ChoreMAP that allows us to define solution maps and display the information extracted from a geographical data base.

Future work can be summarized in the following activities:

- Conduct comparative experimental studies to verify the effectiveness of the proposed method. In particular, such studies aim to measure the ability of the expert users to exploit chorematic maps produced by the system.
- Check chorems can be used as a navigation tool or as a query language and access to the content of geographic databases.

REFERENCES

[1] R. Brunet, ”La carte-modèle et les chorèmes”, Mappemonde 86/4, 1986, pp. 4-6.

[2] R. Brunet, ”Les fondements scientifiques de la chorématique”, In "La démarche chorématique", Centre d'Études Géographiques de l'Univ. Jules Verne, 1993.

[3] I. Cherni, K. Lopez, R. Laurini and S. Faiz, ”ChorML: résumés visuels de bases de données géographiques”, International conference on Spatial Analysis and GEOmatics, Paris, France, 2010, pp. 691-692.

[4] A. Coimbra, ”ChorML: XML Extension for Modeling Visual Summaries of Geographic Databases Based on Chorems”, Master Dissertation, INSA-Lyon, France, 2008.

[5] V. Del Fatto et al., ”Potentialities of Chorems as Visual Summaries of Spatial Databases Contents”, Springer Verlag LNCS, 4781, 2007, pp. 537-548.

[6] V. Del Fatto, ”Visual Summaries of Geographic Databases by Chorems”, thesis INSA de Lyon, 2009.

[7] D. Dorling, ”Cartograms for human geography”, Visualization in Geographical Information Systems, 1994, pp. 85-102.

[8] M. Egenhofer, ”A Formal Definition of Binary Topological Relationships”. In: Foundations of Data Organization and Algorithms, 1989, pp 457-472.

[9] Z. Guo, S. Zhou, Z. Xu, and A. Zhou, ”G2ST: a novel method to transform GML to SVG”. In: 11th ACM international symposium on Advances in Geographic Information Systems, Association for Computing Machinery, 2003, pp 161-168.

[10] J. Han, M. Kamber and A.Tung, ”Spatial clustering methods in data mining: A survey”, In: Geographic Data Mining and Knowledge Discovery. CRC Press 2001, pp. 188-217.

[11] R. Laurini, F. Milleret-Raffort and K. Lopez, ”A Primer of Geographic Databases Based on Chorems”,Springer Verlag LNCS 4278, 2006, pp. 1693-1702.

[12] W. Tobler, ”Les interaction spatiale: solution de W.Tobler, Espace Populations Sociétés”, 1991, pp 467-485.

[13] R. Yann and C. Zanin Tobelem, ”L’Europe dans la régionalisation de l’espace mondiale : étude des flux commerciaux par un modèle d’interaction spatiale”, Géocarrefour, 2009, pp 137-149.

[14] I. Cherni, S. Ouerteni, S. Faiz, S. Servigne, R. Laurini, ”Chorems: A New Tool for Territorial Intelligence”, 29th Urban Data Management Symposium, C. Ellul, S. Zlatanova, M. Rumor, Eds. London, Taylor&Francis, 2013, pp 67-76.

[15] H. Escaith , ”Statistiques du commerce international”, Organisation mondiale du commerce, 2012.