Toward an Efficient Cross-Fertilization of Multiple Information Sources *

Maurice Szmurlo and **Bruno Crémilleux** and **Mauro Gaio** and **Jacques Madelaine**
GREYC — CNRS URA 1526, Université de Caen
Esplanade de la paix, F-14032 Caen Cedex, France.

Abstract. Access and cross-fertilization of multiple information sources is crucial for most Computer Aided Decision Systems. While most of the potential sources may be easily accessed separately by the system, their integration within a homogeneous framework requires a formal representation of their accessibility and semantics. Moreover, confronted to a huge set of database attributes and processing methods, considered as information sources, the end user will be lost, not only because he does not know what sources are available, but also what to do with them. This makes an other level of integration.

This paper presents the state of our current research on the problem of helping the user to discover and use the sources properly. Our strategy is based on the exploration of a teacher-training hyper-document made of maps and their companion textual geo-spacial analysis. It exploits users natural curiosity for discovering the contents of databases and learning how to use the processing methods.

1 Introduction

In geo-marketing, a decider needs to take decisions based on the analysis of one or more aspects of a given phenomenon located in a geographical region. In order to be helped in his reflections, he will build geographical maps showing some aspects he suspects important. Because of the almost infinite amount of phenomena the decider may consider, there is no pre-defined set of maps he could use (except maybe textbook cases). His strategy would therefore be "try and see" if a map enlightens some part of the phenomenon. Then he will do an other trial, and so on, until he has a more precise view of a solution.

The maps the decider builds may be of different nature: they can be quantitative: repartition of the population by age or studies level, average income of the companies, etc., or qualitative, as for example the localization of companies provided the nature of their activity. A short example is described in section 3.1. In both cases the system has to use information and processing methods obtained from various sources and described with different paradigms. The possible sources may be large databases (internal to the companies, or external, as for example national statistics on demography, economy, etc.), geographical databases (GIS-based), results of various processing methods, i.e. data analysis and AI-based exploration methods. The two latter are considered as potential sources since new information is created from the existing ones by computation.

While designing a system which allows the usage described above, we are facing two major problems related the knowledge representation. Integration of multiple information sources into a heterogeneous framework is the first one. This issue will be shortly discussed in section 2. The second problem is related to the representation of the knowledge the user has (or does not have and needs to acquire) about geo-spatial analysis, that is construction of maps and their interpretation.

Most of the geographical CAD systems available now days are able to perfor= m basic integration of sources, but do not provide the user with help on how to choose attributes, how to process them, and how to interpret the produced maps. They are limited to a selection of *all* attributes and methods and let the operator on its own. Unfortunately, unless the user is a specially trained expert, he will not be able to deal with all the information the system provides. Firstly because he doesn't know exactly which attributes are present in the databases, what they represent, and how to access them, and secondly, because he doesn't know precisely which informations should be combined together in order to represent the phenomenon.

This paper presents our current research on this second problem. Our idea is to represent the knowledge on geo-spacial analysis in a dynamic hyper-document containing full examples the decider may come across. The hyper-document (HD for short) will be used on one hand as a tutorial to train the operator and make him discover the contents of the information sources, and on the other hand, as a tool for working by analogy, i.e. by leting the user to create his own maps by modifying dynamically the examples contained in the HD.

The remaining of this paper is organized as follows. In section 2 we hold a short discussion on integration of sources in a homogeneous manner. Section 3.1 explains how the hyper-document is organized and carries out an example of a node of the HD. Finally, section 3.2 shows how the hyper-document may be browsed to discover the contents of the databases and a method to extend this navigation.

2 Integration of multiple sources of information

In order to integrate the sources of information in a homogeneous framework, we need an abstract layer of representation of all the informations that are available. There are two aspects to be represented. The first one addresses the accessibility to the information:

*This research is supported in part by the company Captimark, *32, rue St. Augustin, 75002 Paris, France* and by the European Found FEDER.

- Is an information directly available from a database?
- If yes, in which base?
- Which request is to be sent to the database server, using which protocol, in order to retrieve the information?
- In all cases, what are the necessary resources to perform a request.

The second one addresses the semantics of the informations:

- What does the attribute represent? For example, the attribute "`population`" may represent the male, female, or total population of a given region.
- What is the type of the attributes? Are numerical values absolute or percentages? Are they quantitative or qualitative?
- The two previous informations will allow the system to compute new values whenever a requested attribute does not exist. The system needs therefore the computing rules for these "virtual" attributes. For example, for computing the total population P for a given region, the system may add P_m and P_f, respectively representing the male and female populations for this particular region. However, it is necessary to verify that both P_m and P_f are absolute numerical values: adding percentages together (should) always yields 100%, which is informationless, and adding values of different types is senceless.

A formal and abstract description of the informations will allow consistency checks and will help for integration of new sources. Our work being in its early stage, we did not decide yet how these knowledge and informations should be modelized. This question is therefore not of our concern in this paper. In the recent years, there has been many projects addressing these issues. The interested reader may refer for example, to the research reports on the TSIMMIS project [Chawathe94; Garcia-Molina95] at Stanford University which describes a system for integrating multiple and heterogeneous information sources.

3 The Hyper-Document as the integrator and cross-fertilization facility

As mentioned in the introduction, the user does not have a precise idea of the maps he needs to generate in order to enlighten a phenomenon. The system therefore needs to provide some help both to discover the information and to select the right maps. Unfortunately, due to the very huge amount of informations and to the almost infinite number of aspects the user may want to cartography, it is not possible to provide him with all the maps and interpretations. This is enforced by the fact many databases are integrated within the framework. The help the system may provide is therefore only partial, and the system must include the user in its conception process. That is why we choose to use an hyper-document as a friendly interface. Nevertheless, the navigation in the HD cannot allow the user to browse extensively all the available data. The two following section present the basic principles of the HD and a solution to extend its use for the selection of pertinent data and/or processing methods.

3.1 Principle and example

The hyper-document contains a non-exhaustive list of phenomena that has been analized by experts. Each analysis presents the phenomenon in the main nodes of the HD, as well as the different aspects the experts considers as relevant. Each of the aspects is cartographied on its own page which contains the following informations:

- description of how the current aspect enlightens the phenomenon,
- the map the expert has generated; the map is composed of its graphical representation, legend, title,
- the attributes and processing methods that were used for generating the graphical representation,
- the reasons of the previous choices,
- links to related pages, that is pages presenting the other aspects of the currently analyzed phenomenon.

The hyper-document represents therefore the knowledge of the experts and may be used as a tutorial on geo-spatial analysis. It realizes the user level integration.

The following paragraph presents a simple example of one expert analysis.

Assume the problem is: *The prospective service of a politicaly oriented magazine whishes to know the political opinions of the adult population of a given geographical area in order to send ads.*

- The first aspect that may be interesting to look at is the geographical repartition of the votes between two parties, P_1 and P_2. The produced map shows a great amount of P_1-votes in a few areas. Because we are working with maps, we see that this region are rural.
- The second aspect would be to know what socio-professional categories of the population are located in these areas. The map shows, for example, farmers.
- In oder to validate the hypothesis: "the farmers in these areas have produced P_1-votes", we will build a map representing the correlation between these to data.
- If this last map confirms the hypothesis, the conclusion is that it may be interesting for the magazine to send ads to the farmers of these geographical areas.

All these separate analysis designed by experts contains examples of how informations contained in the available databases are processed in order to obtain a cartography of a given phenomenon. They build up the HD.

3.2 Navigation through hyper-documents to bring out relevant combinations of methods and attributes

The attributes available from the different databases and the processing methods are numerous. Therefore, it does not seem reasonable to present all of them to the user and let him do his own choice. The navigation also permits to restrict the number of attributes and methods presented to the user to the most significant/valuable ones, according the context of the users current work as well as his previous experience with the system.

First, we make the user navigate through the hyper-document. Almost certainly, the user will not find the complete answer to his specific problem in the examples. Therefore we let him the freedom to create new maps, either by modifying the attributes used in the examples, or by using other data processing methods, or both. The new attributes the system offers at this level to the user are very closely related to the ones he's currently working with. For example, if he uses 'population being less than sixteen', the application proposes all the other age categories. On the other hand, the available methods are the ones used in the other examples.

At any time the user can ask the system to validate a given map, that is to save the way the map has been produced in his *profile*. This profile is represented as a parallel hyper-document only accessible to this particular user. This wandering through the hyper-document, permits the user to learn what informations are available in the databases and how to process them: eventually he becomes an 'experienced' user. When an experienced user needs to perform a new task, he may skip this learning phase and directly use his profile, or restart the whole process by browsing the hyper-document.

While this step helps to learn about the processing methods, it does not facilitate the exploration of the attributes present in the databases. The drawback is indeed that the user will mainly use the attributes saved in his profile, without trying to find new ones. Therefore, we also let to the user the freedom to type a request in natural language (or at least to give keywords or "key-sentences"). The system will respond with a list of attributes that it 'believes' would be helpful. For example, for 'industrial repartition' it would propose directly the attribute which represents the industrial repartition in a given geographical area if such information is available, or, more probably, if it does not exists, related attributes that may be used to analyze indirectly this repartition: number of workers, of employees, average size of companies, benefits, etc.. Actually, both cases are of great interest. The first one may be the result of an investigation performed by an institute and really represents the repartition. On the other hand, related attributes, when used either directly or combined with a suitable method, may provide an explanation of the phenomenon and enrich the users reflection.

Analysis of natural language requests is of course a very difficult problem. However, we are working in a very restricted domain, with a limited corpus. In order to facilitate this analysis, the request may be given in semantically typed input fields such as 'date', 'geographical area', 'economical problem', 'social studies', etc. This typing saves a lot of processing work and allows usage of different, specialized, and simple syntaxical and semantical analyzers. This technique has been used with success in a previous project [Enjalbert94; Victorri92]. Because of their types, the expressions are not truely 'natural language' but leave the user the freedom to express larger queries. After the analysis of the request has been performed, a thesaurus [Jing94] defined by the experts will be used in order to retrieve the the relevant attributes unsuspected by the user. Further work is to integrate paradigms from knowledge representation and discovery [Ribeiro95].

4 Conclusion

This article presented the main problems related to the design of a cartography tool integrating multiple and various information sources and we propose a cross-fertilization method based on the usage of an hyper-document. We mainly deal with the selection of valuable database attributes and processing methods in order to respond to the users request. The solution we proposed is fully interactive and exploits users natural curiosity in order to make him discover both the contents of the various databases and the sounded processing methods he may need to use. It is based on a hyper-document that presents already solved problems or studies where the user can navigate to find some clues for his problematic. He can also query the system for attributes in a free format query form, due to the system knowledge of the syntactic and semantic nature of the data and informations. Furthermore, the system guides the end-user in the choice of relevant processing methods of the chosen attributes.

References

[Chawathe94] Chawathe, S. and Garcia-Molina, H. and Hammer, J. and Ireland, K. and Papakonstantinou, Y. and Ullman, J and Widom, J. *The TSIMMIS Project: Integration of Heterogeneous Information Sources*, Department of Computer Science, Stanford University, CA 94305-2140, USA, 1994.

[Enjalbert94] Enjalbert, P. and Victorri, B. *Du langage au modèle*, TAL: Traitement automatique des langues, **35**(1), 1994.

[Garcia-Molina95] Garcia-Molina, H. and Papakonstantinou, Y. and Quass, D. and Rajaraman, A. and Sagiv, Y. and Ullman, J. and Widom, J. *The TSIMMIS Approach to Mediation: Data Models and Languages (Extended Abstract)*, Department of Computer Science, Stanford University, CA 94305-2140, USA, 1995.

[Jing94] Jing, Y. and Croft, W.B. *An association thesaurus for information retrieval*, Technical report 94-17, Dep. of Computer Science, Univ. of Massachusetts, Amherst, 1994.

[Ribeiro95] Ribeiro, J. and Kaufman, K. and Kerschberg, L. *Knowledge discovery from multiple databases*, Proc. IASTED/ISMM Int. Conf. on Intelligent Information Management Systems, June 1995.

[Victorri92] Victorri, B. and Thomazo, L. and Boyreau, G. and Madelaine, L. and Coulon, J-F. and Hanriot, D. and Le Crosnier H. and Girollet, D. *L'antéserveur: une interface intelligente avec l'univers documentaire*, Conf. Int. "Interface entre monde réels et mondes virtuels", **vol** EC2, Montpellier, Février 1992.