Affordable web-based spatio-temporal applications for ad-hoc decisions

Vera Hernández Ernst

Fraunhofer Institute AIS, Schloss Birlinghoven,
53754 Sankt Augustin, Germany
vera.hernandez@ais.fraunhofer.de

Abstract. Nowadays, decision making requires involving complex, multidimensional context information. Therefore, the concept of Spatial Data Warehouses (SDW) offers a suitable solution for fast access of data on different aggregation levels, but the required detailed information and its dimensions are not always available a-priori: (1) user data demand can change unpredictably in the course of the analysis, (2) commercial context data is expensive or needs to be requested on demand from third-party providers like Spatial Data Infrastructures (SDI) and (3) the initial SDW model can change during analysis by adding derived computation results at any level of detail. This paper proposes an approach for building an infrastructure based on open standards to construct and use virtual SDWs dynamically. In particular, services for loading and consolidating new types of spatio-temporal data on demand are discussed. Considering different spatio-temporal models, the paper proposes an implementation solution based on web services.

Keywords: spatial data warehouses, spatial OLAP web services, spatio-temporal decision support

1 Introduction

The analysis of spatio-temporal related data is gaining more and more importance in new application fields through concepts like geo-marketing and spatial data infrastructures (SDI) as well as the increasing presence of mobile devices and (satellite) maps in day-to-day utilities. A hurdle to overcome for many decision makers is, firstly, data acquisition, taking into account the short lifetime and the high costs of current (commercial) spatial data. Secondly, the purchasable data does not always match the analysis requirements precisely enough and must be pre-processed to be used. Thirdly, individual software for multidimensional analysis is expensive and the development of such software is time consuming. Additionally, standard software is often inflexible since it may use proprietary data and/or predefined analysis methods. The use of those systems also requires specific resource skills for collecting, maintaining and analyzing the data, which may be missing on SMEs (small and midsized enterprises).

In addition to the monetary and personnel limitations, there are functional requirements that are not yet solved on commercial multidimensional analysis systems. These are: (1) **federated data sources**: the data model has to be constructed based on distributed data sources, having been collected at levels of detail and on different formats; (2) **multi-level analysis**: the data must be accessible quickly at different levels of detail according to spatial, temporal or any other item grouping aspect; (3) **dynamic structure**: the data structure can be changed during the analysis by adding new data whose structure was previously unknown; (4) **consumable information**: The system should support that data and results be represented properly and be reusable for further purposes; (5) **transparent clearing**: The system should allow the pricing of data and functionality according to the user demands for each analysis

This paper proposes a web-based approach for the dynamical creation and use of multidimensional structures to support SMEs on ad-hoc decisions. Consequently, it is possible for the service providers to quote a transparent price according to the user's demands and choices. After a brief introduction and a motivating application scenario, section 2 will give an overview of related work based on open standards and spatio-temporal models. Section 3 will treat concepts for dynamical virtual SDW while section 4 will outline a solution approach for a SDW-Infrastructure. Finally, section 5 will summarize the state of the work and the future research directions.

1.1 Application scenario

In order to stay on track through this paper, let us choose a specific situation about a real application demand for a small enterprise. For the planning of a GIS congress event, potential locations should be analyzed considering the following aspects:

- ⇨ event locations (e.g. congress centres) and their scheduling, prices and descriptions like terrace, catering, it-infrastructure, number of rooms, etc.;
- ⇨ attractiveness of the location places grouped by the criteria culture, nature, history, sport;
- ⇨ commodities grouped by hotel category;
- ⇨ regional population (socio-demographic data) nearby the available location;
- ⇨ location of specific branches grouped by categories like private enterprises, research institutions, education institutions and administrative bodies. These categories may be subdivided, e.g. private enterprises into data providers, software houses, consulting companies, etc.;
- ⇨ accessibility as distance to next important train station, airport, traffic ways,
- ⇨ distance to other locations offering similar content grouped by themes like IT, geography, environment and traffic,
- ⇨ and data about the locations of former conferences of the same type, to avoid repeating locations.

In order to collect all needed information, the user may choose the basis data sources from different public and private SDI providers, according to the price, the quality and the thematic relevance of the offer. User data should also be loaded to be used as another data source. To choose the best location the user wants to navigate through the data considering different temporal, spatial and thematic detail levels. He may create some maps and charts to explore potential locations interactively. Analysis computations, like catchment areas, may be also inserted into the existing model.

2 Related work

Data Warehouse (DW) and Online Analytical Processing (OLAP) are technologies for modelling and analyzing multidimensional data. Available commercial systems do not support a comfortable involvement of spatial data. The term Spatial OLAP (SOLAP) [14] describes visual tools supporting rapid and easy spatio-temporal exploration of OLAP data. Further related researches offer an extensible guide for development concerning relevant aspects like evolving data [5,13]. [9] proposes a relevant methodology to implement the MADS conceptual modelling approach [12] into Spatial Data Warehouses but does not consider volatile data. Further requirements for the approach are multiple representation methods [3,4,12]. A not often treated issue is the handling of non-numerical measures.

Since most of the approaches found are implemented to be used with commercial tools, interoperable concepts for metadata, communication protocols and mechanisms are needed. A solution could be based on OGC[11] specifications for Web Services. Also relevant is the Common Warehouse Metamodel (CWM OLAP) [8] provided by the Object Management Group (OMG). XML for Analysis (XMLA)[16] offer an open technology that may become a main exchange format for multidimensional data The Framework GeoDWFrame [15] is proposed to integrate OGC and CWM OLAP standards. By the means of GML for Analysis (GMLA)[6], a description for metadata and data can be used by merging GML and XMLA. This approach does not consider transactions, which should be added for volatile data.

A relevant aspect to consider is the cost of third party data. With the development of specifications for authentication, authorization and payment services (Web Pricing and Ordering Services WPOS, OGC discussion paper [11]), there is a solution for an on-demand offer in the foreseeable feature.

3 Building Virtual SDW

One main difficulty of using data coming from SDI's is to structure the data adding semantical and ontological meaning. Indeed, data may refer to identical observations with different denominations, and their combination can lead to imprecision, but this issue will not be considered in this paper. We will assume the data is described through standardized metadata notations, and can be accessed and linked together without complex transformations. Our concept decomposes the complex issues into different facets by using on-line analytical processing (OLAP) models (star- or snowflake-schema) [9].

DWs are mostly constructed on top of historical data, which is extracted from operational systems, transformed and loaded into a previously modelled multidimensional structure. Usually, this is a time and resource consuming process. In order to build the DW "on the fly", the user has to be supported in connecting different data sources and on modelling the hypercubes. Therefore a "conceptual" model will be applied to separate it from the logical and physical realization, based on the (CWM OLAP). The current model will be slightly extended to allow more flexibility on the definition of different kinds of hierarchies [10] and an evolution of the structure during the analysis [5] by storing relationships with attributes [12]. This section will give a formal description as a coarse definition of the conceptual model structure.

3.1 Basic concept definitions

A **Relationship** will be a tuple <Type, [id_from]$^+$, [id_to]$^+$, F, F^{-1}, [Constraint]> where Type characterizes the kind of relationship, [id_from]$^+$ and [id_to]$^+$, are sets of nodes with cardinality 1..n, identifying elements of the structure corresponding to the relationship. The sets may have either the same cardinality, or at least one set must have cardinality 1. F is a mapping function from node id_from to id_to and a confidence factor for this mapping. A confidence factor should give a measure for the quality of the mapping function. If the values are of qualitative nature, a "truth table" should be defined [5]. As consequence, F^{-1} defines the inverse mapping function from id_to to it_from to allow the navigation in both directions. [Constraint] describes a set of rules to be fulfilled to ensure the validity of the relationship. These constraints may be temporal, e.g. for validity times. They can also label a relationship as optional, allowing asymmetric hierarchies [10]. Relationship objects will be modelled as arcs between the OLAP structures (nodes). Thus defining different relationship types a stronger binding to the types of nodes and functions can be defined. :

- ***Aggregation***: (belongs_to) Relationship sorting two levels or two attributes. The cardinality for both sets, [id_from] and [id_to], is 1. id_from indicated the higher aggregated level. This type is used by hierarchy definitions.
- ***Inheritance***: (parent-child) Relationship sorting two classes of different types. [Id_from], with cardinality 1, is the parent, [id_to], with cardinality 1..n, the child object(s).
- ***Association***: Relationship to link two or more members, which can have the same or different types. This relationship may be used to describe topological links.

An **Attribute** is a tuple <Name, Type> used to define a property for a structure element.

A **Dimension** is an abstract structure that allows the user to query the data from different perspectives and is defined by a tuple <Name, [Attribute]> where Name is a unique identifier, and [Attribute] a set of global attributes for the dimension. For each element of [Attribute] there must be at least one Aggregation-relationship for each level. The members of a dimension are levels or hierarchies; last ones are defined by Inheritance-relationships.

Dimensions containing spatial data and those containing temporal data will be treated separately. ***Spatial Dimensions*** must have an attribute ***Representation*** as a tuple <[Feature],[Style], [R_Representation]>, where [Feature] is a set of spatial Features, containing geometrical information and the reference coordinate system, [Style] a set of different Style definitions for drawing the feature geometries and [R_Representation] a set of association Relationships joining Features with their representation. This definition allows multiple representations of dimensions. ***Temporal Dimensions*** must have an attribute ***Time*** as containing a tuple <[TimeMoments],[Format],[R_Timeformating]>, where [TimeMoments] is a set of temporal descriptions, characterized by a referenced calendar, a temporal unit and temporal constraints, like value range. [Format] is a set of predefined output expressions, where each expression may be linked to one TimeMoment by an association relationship (R_Timeformating). The attributes Representation and Time must be defined for each Level that is respectively linked through an Inheritance Relationship to a Spatial or Temporal Dimension.

A **Level** is a dimension member to be used for traversing different dimension stages, defined as a tuple <Name, [Attribute]>. Name identifies the level, [Attribute] is a set of properties describing it. The order of levels inside a dimension to form a hierarchy is defined by aggregation relationships.

An **Observation** represents the information the user wants to analyze at the different levels of detail. It is an Attribute, described by a tuple <Name, Type, [F]>. Usually, for OLAP these attributes are called measures, pointing to the numerical type they should have. Calling them observations allows the use of other data types like categorical, temporal and spatial types. While Name identifies the observation, Type specifies its nature. [F] is a set of grouping functions for the observation. On OLAP an observation may be defined by just one grouping function like average or sum, but for analysis often more than one aggregation measures is relevant[2]. Since on OLAP modelling tools, that can be solved by creating different measures, at a concept and logical level, the user wants to associate the computed values to one observation. At this point it should be considered that by applying some functions the type of an observation may change. For instance an integer value becomes a float by applying an average function and a qualitative value may become quantitative [7]. For this paper we will assume that the functions are type preserving.

A **Cube** expresses an analysis focus as a tuple <Name, [Observation],[Dimension]>, expressing all attributes to analyse as well as the related dimensions. Extending the star- or snowflake schema, a hypercube may contain several cubes defining different analysis tasks, which may share dimensions.

A **DataSource** is a tuple <Type, Name, Metadata, Source> where Type describes the nature of the Source, Name identifies it, Metadata is a link to the metadata description and Source a link to the physical location of the data. A Relationship of type

Instantiation associates a hypercube structure from Type Level, Measure or Observation with a DataSource. The description of the association may be described as Constraint rule, e.g. the path to the attribute on the DataSource.

To visualize the definitions the graphical notations as proposed by [9, 10, 13] can be used.

3.2 Operators

The operations to be performed on the structure can be grouped into four categories: Construction, ETL, Generation, Query and Computation. On this paper we will not give an exact definition of the operators and limit us to enumerate the group members.

(1) **Construction**: operations for creating hypercube structures will be performed involving the Type of the corresponding structure, its identifier and a set of parameters to be filled with the instance values

> ***create_structure, delete_structure, relate*** (constructs a relationship of the given Type), ***change_relationship*** (allows modifying existing relationships by changing the constraints), ***insert_members, delete_member***

(2) **ETL**: operations for associating data to hypercube structures

> ***link_source*** (creates an Instantiation Relationship for a DataSource and a hypercube element), ***insert_object*** (add a new element of a structure object), ***delete_object, update_object*** (modify an element of a structure object according to the given parameter values)

(3) **Generation**: operations to compute the hypercube. Although some aggregation and specialization queries will be performed on demand, some computation intensive associations may be executed and stored a priori, allowing a hybrid olap structure.

> ***validate*** (allows to validate the structure before the hypercube is computed), ***compute*** (executes the functions defined on a relationship and stores the results on a DataSource with the given Name and Type)

(4) **Query**: operations to access and analyze observations and attributes. The operations will be based on the query operations listed by [9]. All the operations will have the input parameters ([Attribute/Observation], [Level members],[parameter]), representing a list of the target information [Attribute/Observation] and a set of level of detail on the target dimensions [Level members]. In case the query depends on some constraints, they may be set as parameter list. The operations should be defined equivalently for the all different data types.

> ***selection*** (Slice), ***projection*** (Dice), ***Cartesian product*** (Tuple construction), ***Drill-up*** and ***drill-down*** (Generalisation, Specification), ***union, difference, intersection, reduction*** (selects values of attributes), ***extension***: (dynamically addition of derived features to a type), ***objectify*** : (attribute transformation via algebra of an object creating a new object and a new relationship), ***fusion***: (group generalization for all group instances), ***merge***: (new type of varying attribute)

Additionally, a set of operations may be applied to present the information to the user.

(5) **Presentation**: Operators to show the queried data on maps, charts, etc, allowing to apply a specific visualisation on attributes or observations. According to the given attribute number and types just suitable representation types can be chosen [2]. A set of parameter can modify the output.

thematic map (cartographical portrayal), ***statistical chart, tabular sheet*** (the values will be presented to the user as table.

4 Implementation approach

The goal of this research is to provide a methodology to allow the demand-driven analysis of multidimensional and hierarchical structured spatio-temporal data based on a service oriented technology. One step is to consolidate research approaches and standardization activities into one single service based approach. Therefore, existing suitable approaches will be integrated and if needed extended, thus validated to be implemented as web services. The idea is to achieve interoperability by exchanging data and requests via standard xml formats and also store temporal data in those format, for instance for uploaded user data, aggregation and specification results etc. The services should be integrated into one logical infrastructure that allows the chaining of services to define particular analytical process steps.

The services can coarsely be structured into the categories:

(1) Conceptual modelling services (CMS): definition and modification of the conceptual model by setting the hypercube structures through implementing construction operations

(2) Content Discovering Services: (CDisS) Catalog Services and harvesting services to find particular content services and to request metadata information

(3) Data linking services: (DLS) Services to allow linking existent data to model structures by realizing ETL operations. Different data sources can also be joined into one concept and user data can be uploaded und transformed to be used as a data source. Other services like geocode, geoparse, gazetteer and coordinate transformation service can be implemented into this group, since they offer methods to support the ETL process.

(4) Data processing services: (DPS) Services to realize analysis computations on the basis of a data structures described by the conceptual model. Here analysis methods like data mining, geostatistical and multicriteria computations can be implemented.

(5) Content access services and transport protocols: (CAS) Services to query data from a data source by a client. The query services implement query operators as filter functions on the data structure. WFS and WCS services are access services as well. XML protocols used to transport the information are also part of this group.

(6) Portrayal services: (PS) they apply presentation operations to query results to be shown by a client. Styles are defined by Styled Layer Description (SLD) files. WMS is a service of this group.

(7) Security and payment services (WPOS): Services for authentication, authorization and payment.

Fig. visualizes a process on which the user goes through different services during the analysis steps, beginning with the structure modelling (1) and ending with the visualisation of results (4). Distributed SDIs can be harvested (2) to link their data into the model (3). Also user data can be loaded into the infrastructure and linked to the model (3). Content Access Services (CAS) may query the external (SDI) and internal data sources, for instance to be used by processing service (DPS). These services compute analysis operations based on the hypercube metadata and the linked data sources. The result of the analysis can finally be delivered to the user by using portrayal services (PS). WPOS services, represented on the diagram as striped filled services, can be concatenated to compute the final analysis prize.

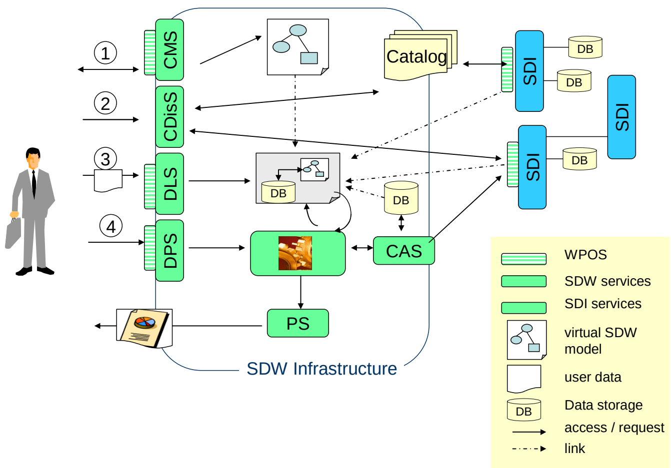

Fig. 1. Process chain and service groupes involved on a SDW infrastructure

5 State of the Work

The present developments aim to create a first testbed joining CommonGIS [1], an interactive tool for spatio-temporal analysis and SpOC; a Spatial OLAP client, which implements the JOLAP[8] API and is capable of discovering OLAP structures and performing OLAP queries and filters. Within CommonGIS, independent variables can be declared as parameters. A wrapper for the CommonGIS structures allows the interpretation of the dimensions as parameters and the dynamic presentation of the query results on the maps. The implemented connectors for OLAP servers use the XMLA[16] protocol to enquire and access data from a Microsoft SQLServer Analysis Service. To support spatial data or dimensions, we used a GIS approach, loading the spatial data into CommonGIS and then joining spatial referenced cube dimension levels such as addresses, postal codes and regions over a common attribute (ID) with the spatial data. By doing this, the data visualizations were updated simultaneously as though navigating through the cubes.

A task support should be offered by adding semantical information to the data processing services. Currently we are extracting analysis methods from CommonGIS to develop decision supporting visualization and analysis services. These may be composed to build workflows of analysis tasks. A semantical description on a service catalog should allow applications to offer alternative visualisations for analogous methods for other data types. In this context, we are developing an abstract language for describing spatial/multidimensional analysis tasks that allows domain-specific analysis steps [2].

5.1 Conclusion and future work

This paper reviewed spatio-temporal modelling approaches and presented some first ideas for a concept and for an implementation framework based on web services to support complex ad-hoc multidimensional and multilevel analysis using open public and private data infrastructures. A main aspect to consider during the implementation will be how to guide the user through the building of models and through the data access. For the scenario the underlying cube will be built and modified on demand and live just for the time of the analysis. For a further approach specialized service providers could be considered, offering bundled data services for application fields, computing basis cubes a priori, which are temporal extended with user data.

The first next research steps will extend the XMLA[16] and GMLA[6,15] protocol specifications to comply the model and data storage requirements presented on this paper. Furthermore, the web service architecture and a selected set of services for model construction, data access and data presentation are intended to be implemented within the framework of the network project “GEOeBizz, web services for adding values to spatial information”, in which two research institutes, three industrial partners (commercial providers for spatial data and software) and a regional SDI provider are involved.

References

1. Andrienko, G., Andrienko, N, Voss, H..: GIS for Everyone: the CommonGIS project and beyond. In: M.Peterson (ed.) Maps and the Internet Elsevier Science, 2003
2. Andrienko, N., Andrienko, G.: Exploratory Analysis of Spatial and Temporal Data A Systematic Approach. Springer-Verlag (2005)
3. Andrienko, N., Andrienko, G.: Intelligent Visualisation and Information Presentation for Civil Crisis Management. In: Suares, J., Markus, B. (eds.), AGILE 2006, 9th AGILE Conference on Geographical Information Science, Proceedings, College of Geoinformatics, Unversity of West Hungary, (2006) 291-298
4. Bédard, Y., Bernier, E.,: Supporting Multiple Representations with Spatial View Management and the Concept of "VUEL"., Joint Workshop on Multi-Scale Representations of Spatial Data, International Society for Photogrammetry and Remote Sensing (2003)
5. Body, M., Miquel, M., Bedard, Y., Tchounikine, A.,: Handling Evolutions in Multidimensional Structures, 19th International Conference on Data Engineering (ICDE'03), (2003)
6. Fidalgo, R. N., Silva. J., Times, V. C., Souza, F.F., Barros, R. S. M.: GMLA: A XML Schema for Integration and Exchange of Mulitidimensional-Geographical Data. In: V Brazilian Symposium on GeoInformatics, Campos do Jordão (2003)
7. Hernández, V., Göhring, W., Voss, A. Hopmann, C. Sustainable Decision Support by the Use of Multi-Level and Multi-Criteria Spatial Analysis on the Nicaragua Development Gateway. In: FIG Working Week 2005 and 8th International Conference on the Global spatial Data Infrastructure, Proceedings, 18-21 April 2005, Cairo, Egypt (2005)
8. JOLAP Expert Group. Java OLAP interface (JOLAP) – proposed final draft, JSR-000069, Technical report, Sun Microsystems Inc., (2003).
9. Malinowski, E.: Concepts and methodological framework for spatio-temporal data warehouse design, Dissertation Thesis, Université Libre de Bruxelles, Belgium, 2003.
10. Malinowski, E., Zimányi, E.: Hierarchies in a Multidimensional Model: From Conceptual Modeling to Logical Representation. In Data & Knowledge Engineering, 2006. To appear.
11. OGC: The Open Geospatial Consortium, Online at http://opengeospatial.org
12. Parent, C., Spaccapietra S., Zimányi E.: Conceptual Modeling for Traditional and Spatio-Temporal Applications. The MADS approach. Springer-Verlag, Berlin Heidelberg New York (2006)
13. Pestana, G., Silva, M., Bédard, Y.: Spatial OLAP Modeling: An Overview Base on Spatial Objects Changing over Time, IEEE 3rd International Conference on Computational Cybernetics (2005).
14. Rivest, S., et al: SOLAP: Merging Business Intelligence with Geospatial Technology for Interactive Spatio-Temporal Exploration and Analysis of Data, Journal of International Society for Photogrammetry and Remote Sensing (ISPRS) "Advances in spatio-temporal analysis and representation, Vol. 60, No. 1, (2005). 17-33
15. Silva, J., Times, V. Fidalgo R. Barros, R.: Towards a Web Service for Geographic and Multidimensional Processing, In Brazilian Symposium on GeoInformatics, Campos do Jordão (2004)
16. XMLA Consortium. XMLA documents. Online at http://www.xmla.org/docs_pub.asp (last access 06/29/2006).