Computers, Environment and Urban Systems 48 (2014) 111–123

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

A new Spatial OLAP approach for the analysis of Volunteered Geographic Information

Sandro Bimonte [a,*], Omar Boucelma [b], Olivier Machabert [b], Sana Sellami [b]

[a] TSCF, Irstea 24 Av. Des Landais, F-63000 Aubiere, France
[b] LSIS, Aix-Marseille University, CNRS, Av. Escadrille Normandie-Niemen, F-13397 Marseille Cedex 13, France

ARTICLE INFO

Article history:
Received 27 December 2013
Received in revised form 22 July 2014
Accepted 22 July 2014
Available online 24 August 2014

Keywords:
Spatial OLAP
Volunteered Geographic Information
Data Quality

ABSTRACT

Volunteered Geographic Information (VGI) has great potential for enhancing analysts' capabilities, assuming that VGI data quality issues, such as credibility and precision, are properly addressed. In this paper, we study the integration of VGI in Spatial OLAP (SOLAP) systems, which allow the integration and analysis of large volumes of good quality data. Using a real-world scenario, we highlight some similarities and differences among these two types of systems. We define a conceptual quality-oriented framework for warehousing and OLAPing VGI data. In particular, to address precision and credibility problems related to VGI data, we propose two new ETL operators: aggregation based on the VGI credibility and a filter based on the historical precision. We also define a new spatio-multidimensional model that provides decision makers with a global description of the quality of the aggregated data. To validate our proposal, we extend the classical relational SOLAP architecture using a standard VGI system.

1. Introduction

Due to the development of new data acquisition technologies, such as sensor networks, mobile devices, and crowdsourcing platforms, geo-referenced data are becoming increasingly available, leading to a wide range of (spatial) data applications, notably Geospatial Business Intelligence (GeoBI) applications. GeoBI systems have been developed to incorporate the powerful analysis capabilities offered by the use of spatial data into decision-making processes. In particular, Spatial Data Warehouse (SDW) and Spatial On-line Analytical Processing (SOLAP) tools (Malinowski & Zimányi, 2008) have been developed to explore and synthesize large volumes of geo-referenced data. These systems rely on closed, well-integrated and structured large data sets. Data are usually extracted from external *official* data sources, cleaned, transformed and loaded into spatial data warehouses using ETL (Extract–Transform–Load) tools according to the spatio-multidimensional model that represents the SOLAP application. The warehoused data are then analyzed using SOLAP clients, which allow data exploration through SOLAP operators using interactive pivot tables and graphical and cartographic displays. Common SOLAP operators include Roll-Up and Drill-Down, which enable climbing hierarchies and aggregating measures, and Spatial Slice, which enables the selection of a subset of warehoused data using topological predicates. SOLAP has been used effectively in several applications, such as health, urban studies, marketing, and the environment (Bernier, Gosselin, Badard, & Bédard, 2009).

Very recently, we witnessed the emergence of collaborative web-based mapping systems, such as Open Street Map,[1] that allow both amateurs (volunteers) and experts to create and share theme-oriented geographic information using various tools, leading to the concept of Volunteered Geographic Information (VGI). Volunteered Geographic Information can be defined as "the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals" (Goodchild, 2007). VGI tools have been used for different applications, e.g., risk monitoring (Goodchild & Glennon, 2010), where recent data, knowledge and participation are required for the decision-making process (Roche, Zimmermann, & Mericskay, 2011). Although proven useful, VGI data suffers from several weaknesses: data quality (*data precision and data credibility*) (Flanagin & Metzger, 2008) (Galindo, Díaz, & Huerta, 2011), scarce metadata, and data that do not always correspond to the "official data" formats (Gouveia & Fonseca, 2008). Moreover, VGI systems do not provide tools for the efficient analysis of historical data (Roche et al., 2011), which is often necessary to understand and explain a geographic phenomenon (Bernier et al., 2009). On the other end, while

* Corresponding author.
E-mail addresses: Sandro.Bimonte@irstea.fr (S. Bimonte), Omar.Boucelma@univ-amu.fr (O. Boucelma), Olivier.Machabert@gmail.com (O. Machabert), Sana.Sellami@lsis.org (S. Sellami).

[1] http://www.openstreetmap.org/.

http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.006
0198-9715/© 2014 Elsevier Ltd. All rights reserved.

proven useful for well-known official data (Vassiliadis, 2009) (Daniel, Casati, Palpanas, & Chayka, 2008), SOLAP warehousing and analytics capabilities are not able to directly handle VGI inputs.

Issues raised by VGI–SOLAP integration have been addressed in (Bimonte, Boucelma, Machabert, & Sellami, 2014). In this paper, we improve our previous work and thoughts (i) by providing a detailed description of the integration framework and approach and (ii) by illustrating the analytics capabilities of the approach by means of an environmental risk case study.

The contributions of this paper are as follows:

(i) We provide a pioneering comparative analysis of SOLAP and VGI systems based on different criteria: users, data, analysis, and architectural issues. Using this analysis, we can identify the theoretical and implementation issues that are highlighted by the integration process.

(ii) We propose a new methodology that considers both the VGI data credibility and precision in the warehousing and online multidimensional analysis phases. We define a new ETL operator that is based on a VGI credibility dimension: *credibility-based aggregation*. We then propose a refinement method which is called the *multidimensional-based precision filter*, for loading VGI data into SDW that compares the VGI data to historical spatial, temporal and thematic warehoused data. Finally, we propose several new SOLAP measures, called *awareness VGI data quality measures*, that allow decision makers to explore VGI and official warehoused data according to the degree of awareness they tolerate for low quality VGI data.

(iii) We propose a UML profile for VGI data and a mapping with an advanced SOLAP UML profile to highlight the generality of our approach and to provide formal definitions for our operators and indicators.

(iv) To demonstrate the feasibility of the new ETL operators and measures, and we propose an ad-hoc architecture (VGOLAP) whereby a VGI tool is strongly connected to a classical Relational SOLAP system using our ETL operators implemented using existing technologies.

The paper is organized as follows. In Section 2, we describe the similarities and differences between the SOLAP and VGI systems. A real environmental case study is introduced in Section 3, and the VGI data-quality-based approach is presented in Section 4. The VGOLAP architecture is detailed in Section 5. Finally, we conclude the study in Section 6.

2. SOLAP and VGI

A Data Warehouse (DW) is a central repository of data integrated from several disparate sources, whose (database) schema derives from a multidimensional model (Kimball, 1996) and whose data are analyzed by On-line Analytical Processing (OLAP) tools. OLAP tools provide the ability to interactively explore multidimensional data, including both detailed and aggregated data. Strategic business decisions are based on the results of these analyses.

Spatial information is often embedded in data, but despite the significance of this information, multidimensional models usually consider only the textual dimension of the data. The integration of valuable spatial data into DWs has led to the development of Spatial OLAP (SOLAP) systems (Bédard, Rivest, & Proulx, 2006), which extend the conventional OLAP model using spatial concepts such as spatial measures and spatial dimensions, in providing support for the representation and storage of spatial data while allowing users to interactively explore and aggregate warehoused data by means of spatial operators. The spatio-multidimensional model that is behind SOLAP systems relies on concepts such as dimensions and facts thatt, respectively, represent the analysis axes and the subjects. Facts are described by numerical attributes (i.e., measures) that are aggregated using classical SQL operations (e.g., SUM, MIN, and AVG) at the different levels of the hierarchies that compose the dimensions.

A typical Spatial Relational OLAP (Spatial ROLAP) system has a three-tier architecture (Bimonte, 2010) (Fig. 1). The first tier is the SDW tier, which stores and manages integrated (spatial) data using a spatial Relational DBMS (Database Management System), whereby the data usually comply with the well-known star and snow-flake schemas, representing facts and dimensions by means of relational tables. The second tier is the SOLAP server tier, which has three main functions: implementing SOLAP operators that compute and handle spatial data cubes, managing the SOLAP schema that establishes a mapping between the spatio-multidimensional model and the SDW schema, and providing the aggregation functions used by the SOLAP operators. Finally, the third tier is the SOLAP client tier, which provides decision makers with interactive visual displays that trigger the SOLAP operators.

Warehoused data are integrated from multiple sources and cleaned using ETL tools, which ensure the desired data quality that is necessary for an effective analysis (Boulil, Bimonte, & Pinet, 2012). SOLAP can be applied to several application domains, such as marketing, archaeology, epidemiology, and environmental risk to cite a few. As an example, environmental data from Siberia (e.g., meteorological, and hydrological data) were analyzed using the "Espla-M" expert system with several data cubes (Penkova, 2012).

As another example, addressing pollution analysis in French departments is presented in (Bimonte, 2010). In that study, the dimensions included (i) the temporal dimension, which was organized into a classical calendar hierarchy (day, month, and year); (ii) a dimension representing pollutants; and (iii) a spatial dimension that represented the French administrative organization of departments and regions. The pollution value was aggregated using the average. Using this SOLAP data cube, users can answer questions such as "What was the average pollution value per pollutant and department in 2000?" or "What is the average pollution value per month per region for inorganic pollutants? (Boulil et al., 2012) (Fig. 2).

Volunteered Geographic Information (VGI) Goodchild, 2007 is a term that is used to define the personal contributions of people to collectively build a geospatial information resource. Citizens' contributions can be viewed as a ladder with several levels; the lowest level considers citizens as simple sensors, while the highest level is a collaboration whereby citizens may be responsible for stating problems and suggesting possible solutions (Haklay, 2010; Murgante, 2013). In the VGI context, a resource could be a geotagged photo, a photo/image-hosting website, such as Flickr, or a data validation system such as Geo-Wiki (Fritz et al., 2011). The term 'neogeography' (Turner, 2006) has been coined to describe a new approach to geography without a geographer (Hudson-Smith, Batty, Crooks, & Milton, 2009) and is related to peoples' activities whereby they create their own maps and geo-tag pictures, movies, and websites. Neogeography can be defined as a new bottom-up approach to geography prompted by users that changes the roles of 'traditional' geographers and 'consumers' of geographical content. VGI has become a reality due to the development of Web 2.0 (Goodchild, 2007), and several VGI systems have been proposed in the literature (e.g., OpenStreetMap and Geo-Wiki). In (Gouveia & Fonseca, 2008), the authors present the main functionalities of VGI systems, which include heterogeneous spatial and multimedia data collecting and processing and data search and visualization using different types of systems (e.g., desktop and mobile platforms). Because they can handle most recent data and because they