BigGIS: A Continuous Refinement Approach to Master Heterogeneity and Uncertainty in Spatio-Temporal Big Data (Vision Paper)

Patrick Wiener[1], Manuel Stein[2], Daniel Seebacher[2], Julian Bruns[3], Matthias Frank[3], Viliam Simko[3], Stefan Zander[3], Jens Nimis[1]

[1]University of Applied Sciences Karlsruhe, Karlsruhe, Germany
{patrick.wiener, jens.nimis}@hs-karlsruhe.de
[2]Data Analysis and Visualization Group, University of Konstanz, Konstanz, Germany
{stein, seebacher}@dbvis.inf.uni-konstanz.de
[3]FZI Research Center for Information Technology, Karlsruhe, Germany
{bruns, frank, simko, zander}@fzi.de

ABSTRACT

Geographic information systems (GIS) are important for decision support based on spatial data. Due to technical and economical progress an ever increasing number of data sources are available leading to a rapidly growing fast and unreliable amount of data that can be beneficial (1) in the approximation of multivariate and causal predictions of future values as well as (2) in robust and proactive decision-making processes. However, today's GIS are not designed for such big data demands and require new methodologies to effectively model uncertainty and generate meaningful knowledge. As a consequence, we introduce *BigGIS*, a predictive and prescriptive spatio-temporal analytics platform, that symbiotically combines big data analytics, semantic web technologies and visual analytics methodologies. We present a novel continuous refinement model and show future challenges as an intermediate result of a collaborative research project into big data methodologies for spatio-temporal analysis and design for a big data enabled GIS.

CCS Concepts

•Information systems → Geographic information systems; •Human-centered computing → Visual analytics; •Software and its engineering → Semantics;

Keywords

knowledge generation, big data analytics, data architecture

1. INTRODUCTION

GIS have long been used to support the decision-making process [2] in many domains like civil planning, environment and nature protection or emergency management. Thereby, geospatial data have always been big data. Petabytes of remotely sensed archival geodata (*volume*) and a rapidly increasing amount of real-time sensor data streams (*velocity*) accelerate the need for big data analytics in order to effectively model and efficiently process complex spatio-temporal problems. In the past, limited access to computing power has been a bottleneck [11]. However, in the era of cloud computing, leveraging cloud-based resources is a widely adopted pattern. In addition, with the advent of big data analytics, performing massively parallel analytical tasks on large-scale data at rest or data in motion is as well becoming a feasible approach shaping the design of today's GIS. Although scaling out enables GIS to tackle the aforementioned big data induced requirements, there are still two major open issues. Firstly, dealing with varying data types across multiple data sources (*variety*) lead to data and schema heterogeneity, e.g., to describe locations such as addresses, relative spatial relationships or different coordinates reference systems [4]. Secondly, modelling the inherent uncertainties in data (*veracity*), e.g., real-world noise and errorneous values due to the nature of the data collecting process. Both being crucial tasks in data management and analytics that directly affect the information retrieval and decision-making quality and moreover the generated knowledge on human-side (*value*). Current approaches mainly address batch and stream analytics in their design that is oftentimes implemented as a closed unified analytical system [16]. While the importance of such systems to efficiently deal with large amount of data is obvious, computers miss the cognition and perception of human analysis to create hidden connections between data and problem domain [14].

In this paper, we present the vision of *BigGIS*, a next generation predictive and prescriptive GIS, that leverages big data analytics, semantic web technologies and visual analytics methodologies. This approach symbiotically combines system-side computation, data storage and semantic web services capabilities with human-side perceptive skills, cognitive reasoning and domain knowledge. We introduce a novel *continuous refinement model* to gradually minimize the real-world noise and dissolve heterogeneity in data and metadata such that the information gain can be maximized. Our con-

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

ISBN 978-1-4503-4589-7/16/10. . . $15.00

DOI: http://dx.doi.org/10.1145/2996913.2996931

tribution lies in (1) an *integrated analytical pipeline* which includes (2) *smart semantic web services*, (3) *domain expert knowledge extraction and generation* as well as (4) *modelling uncertainty* to process high volume, high velocity and high dimensional spatio-temporal data from unreliable and heterogeneous sources. In Section 2, we discuss related work. The platform's design is introduced in Section 3 through the continuous refinement model, while major challenges are presented. Use cases are shown in Section 4. Finally, Section 5 concludes and addresses future work.

2. RELATED WORK

Challenges related to the nature of big data has lead to the evolution of new big data management and analytics architectures embracing big data-aware GIS [12]. Marz proposes the *lambda architecture* [10], a generic, scalable and fault-tolerant data processing system design. By decomposing the problem into three layers, namely batch layer, speed layer, and serving layer, this architecture hybridly combines batch analytics on historic data and stream analytics on streaming data to overcome eachs single weakenesses. Thakur et al. introduce *PlanetSense* [16], a real-time streaming and spatio-temporal analytics platform for gathering geo-spatial intelligence from open source data. Based on the lambda architecture, this platform enriches large volumes of historic data by harvesting real-time data on the fly, e.g., social media, or passive and participatory sensors. While this design allows for adhoc analysis ability during batch runs, the processing logic has to be implemented twice. In a more recent approach, Kreps criticizes the overall complexity of the lambda architecture and presents the *kappa architecture* [9], which simplifies the systems' design by neglecting the batch layer. To replace batch processing, static data is quickly fed through the streaming engine. A representative is *Plasmap*[1], a high performance geo-processing platform that provides a lightweight, interactive query language for high-performance location discovery based on OpenStreetMap. In contrast to BigGIS, both PlanetSense and Plasmap do not apply reasoning on semantic metadata and domain expert knowledge during runtime. However, developments in the field of *semantic web technologies* show the opportunity of adding higher semantic levels to existing frameworks in order to improve their usage in terms of integrating spatio-temporal big data and ease scalability, allowing for reasoning and comprehensive responses [15, 3, 4]. Analyses are often performed in a descriptive, predictive or prescriptive way. While the descriptive analysis visualizes the status quo, predictive and prescriptive analysis focuses on future-oriented planning. As a result, the underlying model and the visualization have to be tightly coupled in order for users to gain knowledge. Users have the possibility to interactively alter a model's parameters according to their knowledge, consequently the visualization adjusts to the model in a feedback-loop. Knowledge generation is one important research area where *visual analytics* is of great use [7, 8], especially when considering uncertainty of heterogeneous spatio-temporal data from various data sources [13]. Jäckle et al. present one possible visualization technique [6] for data and uncertainties of large spatial datasets, which is crucial within use cases where both facets are of importance for decision-making.

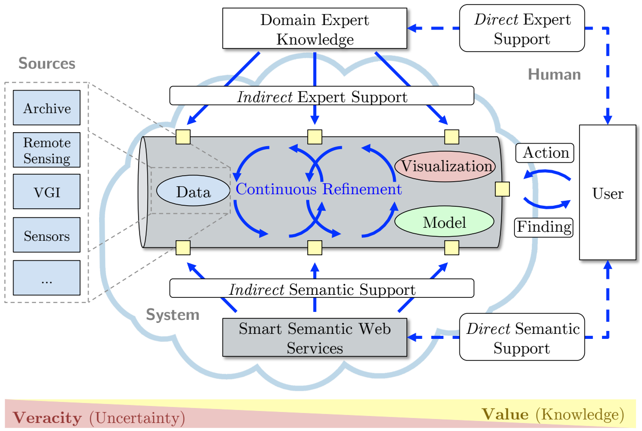

Figure 1: Continuous refinement model in BigGIS.

3. BigGIS PLATFORM

3.1 Continuous Refinement Model in BigGIS

In this section, we briefly describe the continuous refinement model in BigGIS, which extends the knowledge generation model for visual analytics [14]. This will on one hand allow to steadily improve the analysis results, e.g., by updating deployed machine learning models, and on the other hand to build the user's trust in these results by creating awareness of underlying uncertainties and data provenance which is key for providing meaningful predictive and prescriptive decision support in various fields [13]. We consider uncertainty to be reciprocally related to generating new insights and consequently knowledge. Thus, modelling uncertainty is a crucial task in BigGIS. From a high-level perspective, our approach consists of an integrated analytics pipeline which blends big data analytics and semantic web services on system-side with domain expert knowledge on human-side, thereby modelling uncertainty to continuously refine results to generate new knowledge as shown in Figure 1.

3.1.1 Integrated Analytics Pipeline

The analytics pipeline is the core of the continuous refinement model. A key abstraction within this model are specific access points called *refinement gates* that are expressed by a topic-based publish-subscribe pattern (see yellow squares in Figure 1). Refinement gates allow for smart semantic web services, external domain expert knowledge and user interaction to enter the pipeline at arbitrary stages during analyses to continuously improve data management and analyses results, e.g., to support data preparation, to automatically deploy data transformation workflows, to provide domain expert knowledge in order to train machine learning models for pattern detection or to manipulate visualizations.

3.1.2 Smart Semantic Web Services

Locating all available data sources that are relevant for meaningful findings in analytical processes is hard to do when it has to be done manually. Semantic web technologies help to describe data sources using standard vocabularies. Furthermore, reasoning on the logical entailments helps in discovering suitable sources even if they are described differently, providing a two-level support for users through what we call *Linked APIs* and *Cognitive Apps*. The former abstracts

[1] https://plasmap.io/

away the user from manually performing data integration steps to unify heterogeneous data sources by building on appropriate ontologies [4] that support the system (direct semantic support). The latter is a flexible service that is aware of a situational context and capable of sharing it with other services (indirect semantic support).

3.1.3 Domain Expert Knowledge Extraction and Generation

The user is another relevant part in the continuous refinement model who is either provided with additional domain expert knowledge by another person or she herself is the expert in a specific field of application (direct expert knowledge). Overall, we see the continuous refinement process as a knowledge transfer from human to system which is reinforced by smart semantic web services. Thereby, human knowledge is introduced to the system that can contain additional domain specific information and constraints. By doing so, big data analytics can (1) leverage perceptive skills, cognitive reasoning of human analysis to be able to establish hidden connections between data and the problem domain and (2) continuously refine the analyses quality and results. The system intelligently learns from the provided external domain knowledge, such that it can reuse it for future tasks (indirect expert support). Thus, leading to an increasing likelihood of relevant findings by a user during the course of exploration and eventually to generating new knowledge.

3.1.4 Modelling Uncertainty

Uncertainty is inherent in data as well as in models [1]. While this is often obvious in data such as volunteered geographic information (VGI) and participatory sensing data, this holds true for all available data. Models derived from data are only perfectly applicable for the data upon which they are learned. Additionally, domain expert knowledge has some inherent uncertainty as well. Thus, uncertainty constitutes an impediment for refinement. Transparently showing uncertainty would mitigate problems in the first place and would help the user to build up trust. To handle these uncertainties, we express them as *conditional probabilities*. These conditional probabilities allow us to evaluate and model the uncertainty of each data point as well as forecast an analytical model. We apply semantic reasoning on the provenance information of data sources in order to infer a level of uncertainty that can be considered in the analytical processes. We use *bayesian hierarchical models* [1] to be able to cope with the conditional probabilities quantified by the semantic reasoning. The idea behind this is that we can model different parameters by their joint probability distribution. Each parameter can be modelled by hyperparameters, which are again probability distributions. The resulting models are probability distributions as well, which can be used in our continuous refinement model. By doing so, we can model, examine and present the uncertainty at each stage of the process to enable the user of BigGIS to make a well-informed decision.

3.2 Challenges

Data volume and velocity are well-managed requirements through scalable cloud-based big data architectures [10, 9]. Yet, there are still additional big data dimensions, namely variety and veracity of spatio-temporal data, that need to be dealt with in order to generate meaningful knowledge. Based on this, we identify three major challenges.

3.2.1 Varying big data related requirements

The field of application specifies the degrees of big data related requirements. Thus, efficiently managing the complex backend ecosystem for varying requirements is a non-trivial task. We approach this challenge by leveraging Apache Mesos[2] in combination with container technologies such as Docker [3]. In addition, dealing with data and schema heterogeneity and inherent uncertainty is another relevant field of research that BigGIS addresses. Preconditions for meaningful findings in GIS are accurate, consistent and complete data as input for analytical processes. However, as more spatio-temporal data sources emerge the quality of the data is varying as well, especially when considering uncertain data such as VGI. We intend to address this challenge by a smart data integration approach which is based on semantically described data sources extending existing ontologies and data transformation services according to the requirements of different analytical goals [3, 4].

3.2.2 Dimensionality reduction

The continuous refinement model employed in BigGIS aims to provide real-time data processing functionality to minimize the time to deliver insights to the user. Spatio-temporal data, e.g., airborne hyperspectral images, are high-dimensional and models built upon this data have to deal with the curse of dimensionality due to the given real-time constraint. Also, while the presented architecture can handle the challenges of big data, it is not always possible to transfer all the raw data to our pipeline. In the example case of a sensor on an unmanned aerial vehicle, the transfer rate depends on the available bandwidth. BigGIS aims to deal with the challenge of dimensionality reduction for spatio-temporal data, balancing between the robustness of a model and the adaptability to training and input data.

3.2.3 Bias-variance trade-off

The bias-variance trade-off [5] is of particular interest in BigGIS, as the modelling of uncertainty in the continuous refinement model is inherently connected to this. Generally, solving this trade-off optimally is highly depending on the specific use case. Providing the user with sufficient information to reason and generate knowledge under these restrictions is one demanding problem. Here, the challenge lies in the speed of computation, the different level of expertise for each user and the available bandwidth to transfer information back to the user and the analytics pipeline.

4. USE CASES

BigGIS will support decision-making in multiple use cases that require processing of large and heterogeneous spatio-temporal data from unreliable sources. The prototype will be evaluated on three use cases: (1) smart city and health, i.e., heat stress in urban areas, (2) environmental management, i.e., spread of invasive species, (3) emergency management, i.e., identification and dispersion of hazardous gas in chemical accidents. These scenarios represent diverse categories of application domains that each address varying big data related requirements. In brief, an illustrating scenario in the aforementioned emergency management use case is supporting rescue forces in assessing and managing large-scale and

[2] http://mesos.apache.org/

[3] https://www.docker.com/

complex chemical disasters. Providing an in-depth overview of the current situation within a small time frame (velocity) is crucial to prevent exposing the surrounding population to any hazardous substances. Recent developments in the field of mobile robotics allow using in-situ components such as autonomously flying unmanned aerial vehicles equipped with hyperspectral cameras to scan the affected area for hazardous gases producing several gigabytes of raw data per mission (volume). In addition, differing sources (variety), e.g., weather stations, VGI or participatory sensing data, can be integrated in BigGIS, though arriving with high uncertainty (veracity). The combination of those datasets with various other semantically described data sources, helps conducting spatio-temporal statistics. Furthermore, the experts domain knowledge is used to train classifiers in order to automatically classify the hazardous content and identify contaminated areas. Conditional probabilities are computed to forecast the dispersion of the hazardous smoke and visualized in risk maps to highlight potentially endangered areas. Not only are the rescue forces informed about the current situation (descriptive), but also about the risk potential of surrounding areas (predictive), which can be used to automatically alert further public authorities and organizations that would be enabled to perform specifically targeted measures (prescriptive).

5. CONCLUSIONS AND FUTURE WORK

Big geodata will continually grow during the next years. The rapidly increasing distribution and importance of remote sensing data, e.g., from unmanned aerial vehicles, and participatory sensing data as well as the emergence of new data sources lead to more diverse, larger and unreliable data. In this paper, we proposed BigGIS, a next generation predictive and prescriptive GIS, that leverages big data analytics, semantic web technologies and visual analytics methodologies through a novel continuous refinement model. We showed the key architectural elements to master heterogeneity and uncertainty in spatio-temporal big data to generate meaningful knowledge and identified three main challenges. Currently, we are working on an integrated prototype to support each of the presented use cases.

6. ACKNOWLEDGEMENTS

The project BigGIS (reference number: 01IS14012) is funded by the Federal Ministry of Education and Research (BMBF) within the frame of the research programme “Management and Analysis of Big Data” in “ICT 2020 – Research for Innovations”.

7. REFERENCES

[1] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, 2015.

[2] M. D. Crossland, B. E. Wynne, and W. C. Perkins. Spatial Decision Support Systems: An Overview of Technology and a Test of Efficacy. *Decis. Support Syst.*, 14(3):219–235, July 1995.

[3] M. Frank. Integrating Big Spatio-Temporal Data Using Collaborative Semantic Data Management. In *16th Int. Conf. on Web Engineering (ICWE 2016)*.

[4] M. Frank and S. Zander. Smart Web Services for Big Spatio-Temporal Data in Geographical Information Systems. In *ESWC workshop: Services and Applications over Linked APIs and Data (SALAD 2016)*.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2 edition, 2009.

[6] D. Jäckle, H. Senaratne, J. Buchmüller, and D. A. Keim. Integrated Spatial Uncertainty Visualization using Off-screen Aggregation. In E. Bertini and J. C. Roberts, editors, *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015.

[7] D. A. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Information Visualization: Human-Centered Issues and Perspectives*, chapter Visual Analytics: Definition, Process, and Challenges, pages 154–175. Springer, 2008.

[8] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the Information Age: Solving Problems with Visual Analytics*. The Eurographics Association, 2010.

[9] J. Kreps. Questioning the Lambda Architecture. https://www.oreilly.com/ideas/questioning-the-lambda-architecture, July 2014. Accessed: 10 May 2016.

[10] N. Marz and J. Warren. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Manning Publications, 2013.

[11] OGC. Big Processing of Geospatial Data. http://www.opengeospatial.org/blog/1866, August 2013. Accessed: 10 May 2016.

[12] Y. Peng and J. Liangcun. BigGIS: How big data can shape next-generation GIS. In *3rd Int. Conf. on Agro-Geoinformatics (Agro-Geoinformatics 2014)*, pages 1–6. IEEE, August 2014.

[13] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(01):240–249, Jan 2016.

[14] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge Generation Model for Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, December 2014.

[15] V. Tanasescu, A. Gugliotta, J. Domingue, R. Davies, L. Gutiérrez-Villarías, M. Rowlatt, M. Richardson, and S. Stinčić. A Semantic Web Services GIS Based Emergency Management Application. In *Proc. of the 5th Int. Conf. on The Semantic Web*, pages 959–966. Springer, 2006.

[16] G. S. Thakur, B. L. Bhaduri, J. O. Piburn, K. M. Sims, R. N. Stewart, and M. L. Urban. PlanetSense: A Real-time Streaming and Spatio-temporal Analytics Platform for Gathering Geo-spatial Intelligence from Open Source Data. In *Proc. of the 23rd SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems*, pages 11:1–11:4. ACM, 2015.