ADVISU: Interactive Visualization of Anomalies and Dependencies from Massive Scientific Datasets

Noël Novelli[*], Laure Berti-Équille[*,**], Christophe Hurter[*,***]

[*] LIF
163 avenue de Luminy
F-13288 Marseille Cedex 9, France
noel.novelli@lif.univ-mrs.fr
http://pageperso.lif.univ-mrs.fr/ noel.novelli/
[**]IRD, ESPACE DEV
Maison de la Télédétection
500, Rue J.F. Breton, F-34093 Montpellier cedex 05, France
laure.berti@ird.fr
http://www.cerege.fr/Laure.Berti-Equille/
[***]DGAC/DSNA/DTI/R&D/MTC
7 Avenue Edouard Belin, F-31055 Toulouse cedex, France
christophe.hurter@aviation-civile.gouv.fr
http://perso.tls.cena.fr/hurter/

Abstract. In this demo, we present ADVISU (*Anomaly and Dependency VISUalization*), a powerful interactive system for visual analytics from massive datasets. ADVISU efficiently computes different types of dependencies (FDs, CFDs) and detects data anomalies from databases of large size, *i.e.*, up to several thousands of attributes and millions of records. Real-time and scalable computational methods have been implemented in ADVISU to ensure interactivity and the demonstration is intended to show how these methods scale up for real-world massive scientific datasets in astrophysical and oceanographic application domains. ADVISU provides the users informative and interactive graphical interfaces for visualizing data dependencies and anomalies. It enables the analysis to be refined interactively while recomputing the dependencies and anomalies in user selected subspaces with good performance.

1 Introduction

Scientific and corporate applications are continuously producing data at ever-increasing rates. This explosion of data is overwhelming our capabilities to explore, analyze, hypothesize, and thus fully interpret some underlying phenomena of interest (traffic monitoring, environment pollution, global warming, exoplanet identification, etc.). These tasks are becoming even more challenging as we have to detect and handle various and intricate types of anomalies before further advanced analysis. Visualization is one of the linchpins for solving these challenges. To facilitate the analysis of large databases, a combination of effective visual human-computer interfaces and powerful computational analytics have to be proposed for detecting data anomalies and dependencies and offering appropriate summarization artifacts. In this demo, we are particularly interested in showing anomaly detection and recomputation of FDs and CFDs from very large datasets and also from data subspaces that have been interactively selected by the user. Interestingly, we observe that some FDs and CFDs may rely on erroneous data for which we provide appropriate visual representations.

2 Challenges of Visual Analytics

The main scientific challenges addressed by ADVISU are the following:
Computational challenges. Ultimately, ADVISU aims to visually present the users with all the relationships, dependencies and patterns of interest from very large datasets. The types of data dependencies which we are primarily interested in are the exact and approximate functional dependencies (FDs and AFDs, respectively) and Conditional Functional Dependencies (CFDs). Current methods (such as TANE (Huhtala et al., 1999), FUN (Novelli and Cicchetti, 2001) for FDs, and (Bohannon et al., 2007) for CFDs) are not capable of analyzing such large amounts of data. Although all these approaches optimize the consumption of CPU power and memory, they perform all computations in memory. In our context, data loading usually exceeds the memory capacity. It is therefore essential to redesign xFDs discovery and anomaly detection algorithms to take into account the data transfers between main memory (RAM) and secondary memory (hard disks) in the algorithms. Because disk access is significantly more expensive than access to RAM, we propose strategies to minimize computation time by: *i)* storing temporary computation results on disk and then retrieving them when necessary, *ii)* dropping the results of some computations or re-computing them if necessary using the data already present in RAM, or finally, *iii)* approximating the final results from sketches or aggregates computed from on-the-fly sampling and data mining, and by refining the results gradually. Similarly, these computational challenges have to be addressed for anomaly detection (Kriegel et al., 2010).
Technical challenges. The amount of data we handle is such that the computation must be distributed on several computing units. In particular, the computations will be performed by the CPU (multi-threads) and visualization will be associated with the GPU. The ADVISU display offers two main advantages: *(i)* multi-point-of-view is based on various criteria for visual zooming in/out, *(ii)* interactivity in massive data exploration requires that the computation methods for building data representations have to be extremely efficient, on-the-fly and on-demand.

3 The Design of ADVISU

We would like our tool to be used by users who would benefit from *ad hoc* interactive visual analytics while browsing the dataset. Users are fairly non-technical, so the interface has been designed to be straightforward. We assumed no prior knowledge of database or data mining technology.

Designing the visual analytics interface. As discussed previously, the data space is very large. To enable the users to effectively find relevant and valid dependencies from such massive data, we propose a design that narrows the search space and the corresponding visualization universes to be anchored to a specific, user-specified seed subspace. In this demo, we focus on helping the user *i)* understand the space of data analytics through appropriate representations of data dependencies, *ii)* understand that these xFDs may be discovered on-the-fly both from erroneous and correct data, and *iii)* handle the differences moving from one space to another.

Finding dependencies and anomalies. To extract dependencies, we have extended FUN and CFUN approaches (Novelli and Cicchetti, 2001). These approaches can discover all valid (exact, approximate or conditional) functional dependencies (xFDs) using the same theoretical framework (freeset, closure and quasi-closure). The associate algorithms are and the set of valid xFDs discovered represents the canonical cover of xFDs (the set of the minimal and non-trivial xFDs with the right hand side reduced to one attribute). The pruning rules are based on the minimality of left hand side of xFDs that must not contain xFDs (freeset). If a candidate is a non-freeset, it is removed and its supersets will not be generated. Outlier detection in ADVISU is based on various multivariate techniques presented in (Kriegel et al., 2010) and the detection of the other types of anomaly (e.g., duplicate records, inconsistencies, and missing values) is based on state-of-the art methods reported in (Berti-Equille et al., 2011) and *ad hoc* constraints.

4 Implementation and Demonstration

Implementation. All algorithms were implemented using the language C++ with QT for HCI. An executable file can be generated with MS Visual C++ or GNU g++ compilers. Experiments were performed on an Intel Core 2 Quad Q9550/2.83 GHz with 8GB running on Windows 7. For FD discovery, we have upgraded FUN and CFUN approaches to compute partitions and partitions products with disk accesses. To reduce the memory requirement, we strip partitions when the number of elements in equivalence class is less than a given threshold. This produces an approximate cover of valid FDs over a considered relation. The user can then refine the exploration of FDs interactively.

Datasets. The following real-world datasets are used for the demonstration scenario: Corot-exodat for exoplanet discovery (`http://cesam.oamp.fr/exodat/`) and Cosmos databases (`http://cencos.oamp.fr/hstcosmos/`), and oceanic data (`http://www.psmsl.org/data/`) for long term sea level change information from tide gauges and bottom pressure recorders.

Demonstration Scenario. The visualization and exploration of these datasets are challenging, regarding the limited number of pixels available on the screen, and the computation time to interact with the data. It is especially the case when dealing with datasets containing uncertainty, missing and erroneous values. To face these challenges, we implemented interactive features in our exploration tool. First, the user can interactively filter the initial data to display only subset of it. The user can then adjust the filtering parameters to add or remove data, select dependencies based on specific fields of the dataset or anomaly metrics. This technique helps the user to reduce the displayed data and thus the view cluttering. Second, we added interaction techniques based on *mole view* deformation (e.g., see video of (Hurter et al., 2011)) and *focus plus context* techniques. The basic idea is to enable the user to see the object of primary interest presented in full details while at the same time getting an overview of deformed information. For instance, the user can display extracted FDs, and have in the view surrounding of the view anomaly metrics. Third, the user can directly interact with the data. For instance, the user will visualize not only discovered FDs and CFDs but also the validity of these FDs will be displayed since they may be computed from erroneous values.

References

Berti-Equille, L., T. Dasu, and D. Srivastava (2011). Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In *In Proc. of ICDE 2011*, pp. 733–744.

Bohannon, P., W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis (2007). Conditional functional dependencies for data cleaning. In *Proc. of ICDE'07*, pp. 746–755.

Huhtala, Y., J. Karkkainen, P. Porkka, and H. Toivonen (1999). Tane : An efficient algorithm for discovering functional and approximate dependencies. *Comp. J. 42*(2), 100–111.

Hurter, C., A. Telea, and O. Ersoy (December, 2011). Moleview: An attribute and structure-based semantic lens for large element-based plots. In *IEEE Transactions on Visualization and Computer Graphics*, pp. 2600–2609. `http://perso.tls.cena.fr/hurter/papers/InfoVis2011MoleView.wmv`.

Kriegel, H.-P., P. KrÃűger, and A. Zimek (2010). Outlier detection techniques, tutorial. In *SIGKDD 2010*.

Novelli, N. and R. Cicchetti (2001). Functional and embedded dependency inference: a data mining point of view. *Information Systems (IS) 26*, 477–506.

Résumé

Dans cette démonstration, nous présentons ADVISU (*Visualisation de dépendances et d'anomalies*), un système interactif performant pour l'analyse exploratoire visuelle de grandes masses données. ADVISU calcule efficacement différents types de dépendances (DFs, CDFs), détecte des anomalies dans des bases de données de très grande taille, *i.e.*, jusqu'à plusieurs milliers d'attributs et plusieurs millions d'enregistrements. Des méthodes de calcul temps-réel ont été implémentées dans ADVISU et la démonstration a pour but de montrer comment ces méthodes passent à l'échelle pour de très gros jeux de données scientifiques réels dans les domaines de l'astrophysique et de l'océanographie. ADVISU fournit aux utilisateurs des interfaces graphiques informatives et interactives pour visualiser à la fois les dépendances et les anomalies identifiées dans les jeux de données. Il leur permet de raffiner les résultats d'analyse de façon interactive en recalculant les dépendances et les anomalies avec de bonnes performances sur des sous-espaces du jeu de données sélectionnés par l'utilisateur.