What's Up in Business Intelligence? A Contextual and Knowledge-Based Perspective

Marie-Aude Aufaure

Ecole Centrale Paris, MAS Laboratory, Chatenay-Malabry, France
Marie-Aude.Aufaure@ecp.fr

Abstract. The explosive growth in the amount of data poses challenges in analyzing large data sets and retrieving relevant information in real-time. This issue has dramatically increased the need for tools that effectively provide users with means of identifying and understanding relevant information. Business Intelligence (BI) promises the capability of collecting and analyzing internal and external data to generate knowledge and value, providing decision support at the strategic, tactical, and operational levels. Business Intelligence is now impacted by the Big Data phenomena and the evolution of society and users, and needs to take into account high-level semantics, reasoning about unstructured and structured data, and to provide a simplified access and better understanding of data. This paper will depict five years research of our academic chair in Business Intelligence from the data level to the user level, mainly focusing on the conceptual and knowledge level.

Keywords: Business Intelligence, Semantic Technologies, Content Analytics.

1 Introduction

Business Intelligence main objective is to transform data into knowledge for a better decision-making process. We have now entered the era of knowledge. Ubiquitous computing as well as the constant growth of data and information leads to new ways of interaction. Users manipulate unstructured data – documents, emails, social networks, contacts – as well as structured data. They also want more and more interactivity, flexibility, dynamicity. Users expect immediate feedback, and want to find information rather than merely look for it. Decision-making is more and more rapid and we need to automate more decisions. Moreover, the company tends to be organized in a collaborative way, called enterprise 2.0. All these evolutions induce challenging research topics for Business Intelligence, such as providing efficient mechanisms for a unified access and model to both structured and unstructured data. Extracting value from all these data, a crucial advantage in an hypercompetitive market, requires content analytics. In order to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data, the use of data visualization techniques becomes critical. This leads to visual analytics, which combines automated analysis

W. Ng, V.C. Storey, and J. Trujillo (Eds.): ER 2013, LNCS 8217, pp. 9–18, 2013.

techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [1]. Business Intelligence is also impacted by big data, and need to take account of the heterogeneity and the volume of data sources as well as the need of response in real-time for extracting value from trusted data. Finally, Business Intelligence will integrate collaborative and social software. This is known as Business Intelligence 2.0 [2], which combines BI with elements from both Web 2.0 and the Semantic Web.

The organization of this paper is the following: Section 2 first overviews the evolution of Business Intelligence. Section 3 is dedicated to the presentation of research work we conducted in the academic chair in Business Intelligence during five years. Finally, Section 4 describes our future work and related challenges.

2 From Classical to Modern Business Intelligence

Business Intelligence (BI) refers to a set of tools and methods dedicated to collecting, representing and analyzing data to support decision-making in enterprises. BI is defined as the ability for an organization to take all input data and convert them into knowledge, ultimately, providing the right information to the right people at the right time via the right channel. During the two last decades, numerous tools have been designed to make available a huge amount of corporate data for non-expert users. Business Intelligence is a mature technology, widely adapted, but faces new challenges such as incorporating unstructured data into analytics. These challenges are induced by constantly growing amounts of available data. A key issue is the ability to analyze in real-time these data, taking their meaning into account. This amount of information generated and maintained by information systems and their users leads to the increasingly important concern of information overload. Personalized systems and user modeling [3] have thus emerged to help provide more relevant information and services to the user. The complexity of BI tools and their interface is a barrier for their adoption. Thus, information visualization and dynamic interaction techniques are needed for a better use of such tools. Semantic technologies [4][5] focus on the meaning of data and are capable of dealing with both unstructured and structured data. Having the meaning of data and a reasoning mechanism may assist a user during her analysis task.

Traditional BI systems can be extended with semantic technologies to capture the ***meaning*** of data and new ways of interacting with data, intuitive and ***dynamic***. The vision of the CUBIST[1] project is to extend the ETL process to both structured and unstructured data, to store data in a triple store and to provide user-friendly visual analytics capabilities.

Fig. 1 (a) depicts the classical architecture of a Business Intelligence system in which data sources are structured and loaded in a data warehouse. Users can interact with restricted queries producing a static dashboard. In the CUBIST vision (Fig. 1 (b)), data sources are heterogeneous, both structured and unstructured and semantically stored in a triple store. Users can interact with flexible queries and dynamically perform visual analytics. Fig. 1 (c) depicts the most recent trend for BI taking into

[1] CUBIST EU FP7 project: `http://www.cubist-project.eu/index.php?id=378`

account ***streams and semantics***. With the exponential growth of sensor networks, web logs, social networks and interconnected application components, large collections of data are continuously generated with high speed. These data are called "*data streams*": there is no limit on the total volume of data and there is no control over the order in which data arrive. These heterogeneous data streams [6] are produced in real time and consequently, should be processed on the fly. Then, they are maintained, interpreted and aggregated in the purpose of reusing their semantics and recommending relevant alerts to the targeted stakeholders in order to react to interesting phenomena occurring in the input streams. A precious decision-making value can be enhanced through the semantic analysis of data streams, especially while crossing them with other information sources. Real-time BI can be linked to Semantic BI, or can represent an independent evolution. In what follows, we consider that real-time BI integrates semantic technologies for being able to capture the meaning of large amounts of data.

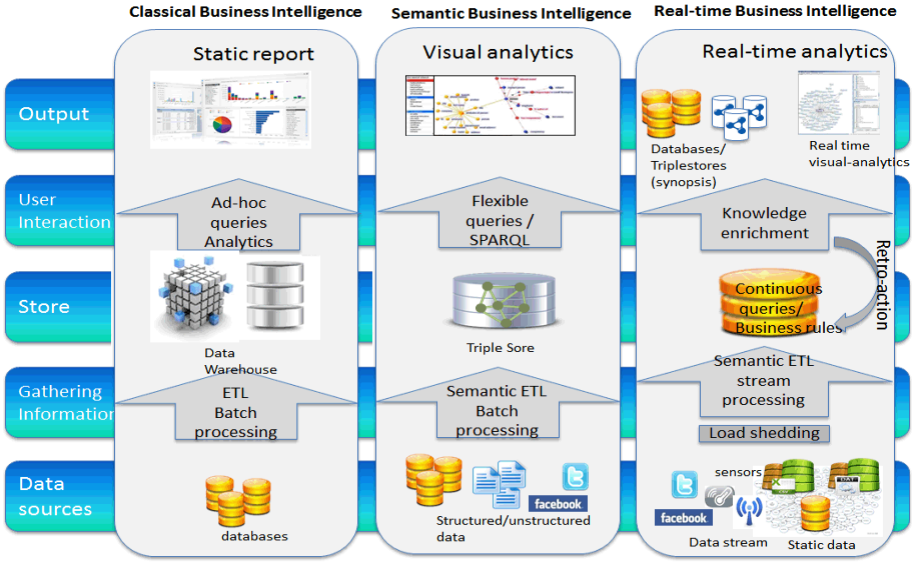

Fig. 1. Classical BI (a), Semantic BI (b), Real-Time BI (c)

In a Real-Time BI platform, multiple heterogeneous data sources can be connected, and data can be static or dynamic. The static data comes from standard databases or from open data, and does not change or in a minor way. Dynamic data comes as a stream, in a semantic format (RDF for example) or not (raw data). After their capture, data streams and static data are submitted to a set of semantic filters designed to achieve some specific business process. To manage infinite real-time data stream, the platform has to provide the ability to create persistent continuous queries, which allow users to receive new results when they become available.

Semantic filters are used to interconnect streams from various sources and to perform reasoning on these interconnected streams, possibly merged with static data (knowledge databases or ontologies). Semantic filters can also include summarizing

operators that extract subset of data in different form (uniform random samples, clusters, patterns, etc.). Load shedding [7] drops excess load by identifying and discarding the relatively less important data. The results of semantic filtering can be for example an alert, a new data stream, enrichment of ontology, feeding a dashboard, etc. Results can also cause feedback (stream connected back into the platform) in order to improve treatment or to add a context to the current treatments.

3 Semantic and Context for Business Intelligence

During five years, we have developed research and tools integrating structured and unstructured data for a more user-friendly decision-making process[2]. The objective was to develop a knowledge layer allowing the end user to easily get the meaning of vast amounts of data, and visual analytics and querying tools for a user-friendly access to data. Fig. 2 summarizes the research work we developed:

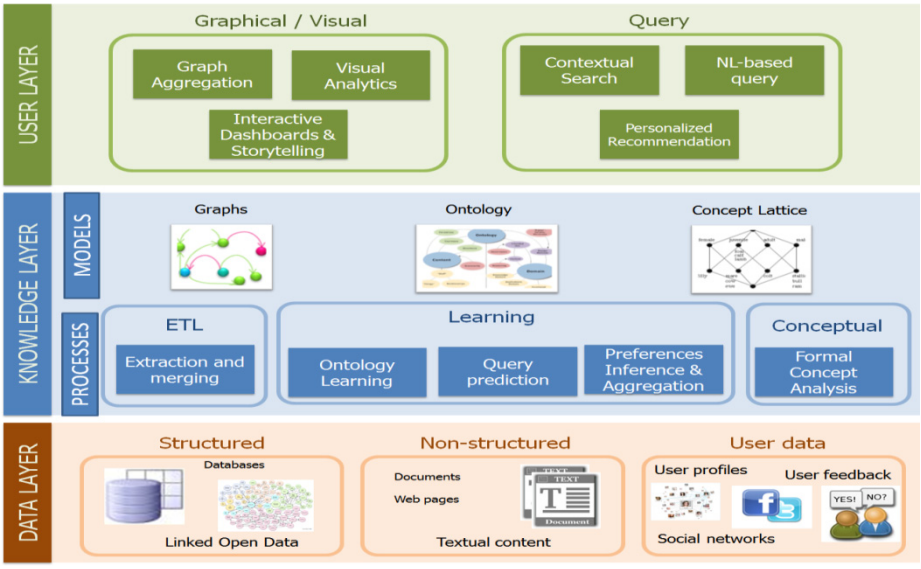

Fig. 2. A vision of Semantic and Contextual Business Intelligence

3.1 Data Layer and Models

The data layer is composed of a set of heterogeneous data sources: structured ones (databases and Linked Open Data), unstructured content (documents, web pages) and user data (user profiles and feedback, social networks). These data sources are structured by the mean of a knowledge layer. Graphs can be seen as a natural way for representing both structured and unstructured information. User's situation and interactions can be represented by a graph, and operators can be applied on a particular

[2] This work has been funded by SAP and EU FP7 projects CUBIST and PARLANCE

graph in response to business events. Moreover, one can benefit from graph operators and algorithms (shortest path, graph transformation) as well as social networks metrics (centrality). Semantic technologies, and particularly ontologies, can be useful for many applications. This model is central in our research work and used in many processes we defined. Formal Concept Analysis (FCA) captures hitherto undiscovered patterns in data. Formal concepts (also called concepts) are formalized as groups of objects associated with groups of attributes. Hierarchical relationships between these groups are formed and visualized by Galois Lattices in the form of a Hasse Diagram [8]. Information organized in this way has a close correlation to human perception but the challenging issue is to provide intuitive visualizations of Galois Lattices.

3.2 Knowledge Layer Processes

The first one is an ***ETL process*** (Extraction, Transformation, Load) extracting and merging graphs from relational databases, modeled using a complex graph structure (like typed attributed graphs). This transformation of relational databases into a graph allows the user to discover hidden relationships between objects [9]. However, this process still induces challenges mainly related to names resolution and scalability issues for merging graphs extracted from different databases.

The next set of processes is related to ***learning***: ontology learning, query prediction based on users' OLAP sessions, and preferences inference and aggregation.

Ontology learning [10] is used to dynamically build a knowledge base, composed of ontology modules, from web abstracts and users' queries [11]. This process has been designed in an automatic and domain-independent way, exploiting unsupervised techniques and the web as a social scale learning source. We exploited these modular ontologies for semantic search, combining them with case-based reasoning. A case is defined by a set of similar queries associated with its relevant results. The case base is used both for ontology module learning and for contextualizing the search process. Module-based similarity is used to retrieve similar cases and to provide end users with alternative documents recommendations. Finally, with the rapid growth of structured data on the Web, referred as the Linked Open Data (LOD) including over 31 billion triples interlinked by around 504 million links, we have extended our approach and exploited DBpedia as a way to bootstrap the learning of linguistic patterns from unstructured data on the Web [12]. Subsequently the self-learned patterns are used for the extraction of new entities and ontology enrichment. In order to do this, we applied deep shallow syntactic analysis by using grammatical dependency analysis on Web snippets provided iteratively by a search engine according to automatically generated queries. Open research issues still remain. Among them, we can cite scalability issues and how to automatically find relevant data sources and ontologies for open domain or specific applications.

The ***query prediction*** process is related to infer the next possible OLAP query based on recent analytical sessions. In Business Intelligence systems, users interact with data warehouses by formulating OLAP queries aimed at exploring multidimensional data cubes. Being able to predict the most likely next queries would provide a

way to recommend interesting queries to users on the one hand, and could improve the efficiency of OLAP sessions on the other. In particular, query recommendation [13] would proactively guide users in data exploration and improve the quality of their interactive experience. Our framework for predicting the most likely next query and recommend it to the user relies on a probabilistic user behavior model built by analyzing previous OLAP sessions and exploiting a query similarity metric [14].

Preferences inference and aggregation are related to user model learning, critique-based mechanism for query refinement and sentiment analysis. The first two processes are related to spoken-dialogue in the context of the PARLANCE[3] European project. Many current spoken dialogue systems for search are domain-specific and do not take into account the preferences and interests of the user. In order to provide a more personalized answer tailored to the user needs, we propose a spoken dialogue system where user interests are expressed as scores in modular ontologies, each ontology module corresponding to a search domain. This approach allows for a dynamic and evolving representation of user interests. Concepts and attributes in hierarchical ontology modules are associated with weight vectors expressing the interest or disinterest of a user on different levels of granularity. A key challenge for personalized mobile search is to tailor the system answers to the specific user and his current contextual situation. In particular, it is essential for recommender systems to perform preference adjustments based on user feedback, in order to modify the actual user behavior. However, regardless of the importance of user preference adjustments, it is not a trivial task. To tackle this challenge, we developed a preference-enabled querying mechanism for personalized mobile search by adjusting user preferences according to user's critiques and refining the queries with respect to the adjusted preferences [15].

Preferences aggregation deals with sentiment analysis [16]. Considering the impressive amount of unmediated opinions expressed by users in social network environments, we analyzed this data with the goal of automatically detecting the polarity of their opinions and perform recommendations. For this, we presented an approach to feature-level sentiment detection that integrates natural language processing with statistical techniques, in order to extract users' opinions about specific features of products and services from user-generated reviews [17].

The last process is related to ***Formal Concept analysis*** (FCA). We used FCA in the context of social network analysis [18], for classifying tweets on specific topics. We also use FCA to flexibly and efficiently build user communities. This entails a novel approach to represent dynamically evolving user preferences and interests. By analyzing the dialogue history of the user, interests are inferred and ontology modules for different domains are annotated with scores. The interests are used to perform formal concept analysis and to construct ad-hoc communities of users sharing similar interests, allowing a form of social search. By collaborative filtering we can share and recommend possibly interesting information and additional communities to users.

---

[3] PARLANCE EU FP7 project: `https://sites.google.com/site/parlanceprojectofficial/`

3.3 User Layer

The user layer is structured in two ways: by having graphical/visual support and by providing textual or formulation support.

Our research in ***information visualization*** has so far focused on 2 directions: (i) improving user experience when using BI dashboards and (ii) developing new visualization and interaction techniques for exploring data that are applicable to multiple domains (e.g. BI, intelligence analysis, social sciences). The main contributions from this work have been new ways to visualize data changes on charts, a new context aware annotation model for dashboards, as well as a set of domain independent interaction and visualization techniques for different data (usually in graph form).

A gap in the application of FCA to Business Intelligence concerns visual analytics. In FCA, the hierarchical relationships between concepts are traditionally displayed as a line diagram representation of the lattice. The concept lattice visualization can be greatly enhanced by visual analytics features and interlinked with best practices from known BI visualizations. A challenge is to manage and navigate the complex concept interrelationships, by condensing and clustering the results, and by sub-dividing and filtering data.

Graphs, and more specifically social networks, can have a huge size making them difficult to analyze and interpret. Producing meaningful summaries from complex graphs, taking into account multiple relations between nodes and various attributes in nodes, is a necessary step. We extended an aggregation algorithm, and defined two new aggregation criteria to improve the quality of the results and experimented them on various graphs.

As analysts continue to work with increasingly large data sets, data visualization has become an incredibly important asset both during sense making analysis, and when communicating findings to other analysts, decision makers or to a broader public. Individually and collectively, stories help us make sense of our past and reason about the future. Given the importance of storytelling in different steps of the analysis process it is clear there is a need to enhance visual analysis tools with storytelling support. We followed user-centered design approach to implement a storytelling prototype incorporated in an existing visual analysis dashboard.

Information access can also be managed through a ***textual or formulation support*** like NLP-based queries or for managing query reformulation and personalized access. A very promising use-case for Question-Answering (Q&A) over structured data is Business Intelligence. Understanding and converting an end-user's natural language input to a valid structured query in an ad-hoc fashion is still a challenging issue. Following this direction, we have proposed a framework for Q&A systems able to define a mapping between recognized semantics of a user's questions to a structured query model that can be executed on arbitrary data sources [19]. It is based on popular standards like RDF and SparQL and is therefore very easy to adapt to other domains or use cases. Personalization and recommendation techniques are useful to suggest data warehouse queries and help an analyst pursue its exploration. We defined a personalized query expansion component which suggests measures and dimensions to iteratively build consistent queries over a data warehouse [20]. Our approach leverages (a)

semantics defined in multi-dimensional domain models, (b) collaborative usage statistics derived from existing repositories of Business Intelligence documents like dashboards and reports and (c) preferences defined in a user profile.

4 Future Research Axis

Nowadays, in order to efficiently and effectively access the vast stream of information available on and off-line, the users and/or enterprises need to resort to flexible and efficient platforms or integration services.

The state of the art, however, presents various dilemmas between data, scale and temporary constraints; on the one hand, for example, common platforms and/or application which try to exploit information from available data stream (whatever they are) tend to still rely on standard keyword-based analysis and queries, and pure matching algorithms (using word frequencies, topics recentness, documents authority and/or thesauri) to find suitable information relevant to their needs. The result is that those platforms are not able to lead users into an intuitive exploration of large data streams because of their cumbersome presentations of the results (e.g. large lists of entries) and, above all, their missing interpretation of the data.

On the other hand, since data arrives continuously, fast one-pass algorithms are imperative for real-time processing of streams. Many solutions have been developed recently for processing data streams, however, none of them takes into account the semantic knowledge of data (stream reasoning). Moreover, the existing platforms do not provide any mechanism to retrieve other relevant information associated to those contents and, even if user feedback methods are proposed, it is really hard to detect the requested information because of inexperience and common lack of familiarity with the proposed visual tools.

Ontologies and, more generally, Linked Data can be now exploited due to its rich collection of structured information, as well as the “deep Web” of database-backed contents. The richness of these semantic data offers promising opportunities for identifying and disambiguating entities from one or multiple sources.

For this, the aim of a successful and effective platform which goal is to process and analyze dynamic stream of data should be the recognition and identification not only of those entries that are relevant to some high-level user need, but also the identification of the entities semantically related to this need and their disambiguation with respect to alternative similar entities.

Moreover, considering that one of the key aspects that has led to the popular success of user-generated content is the possibility to express (and read) unmediated individual opinions, we believe that a successful platform should try to integrate this knowledge for improving its performance. In fact, nowadays, users are currently expressing their opinions and interests on a wide range of entities, products and services. The importance of this data in search and decision making processes is therefore evident. In fact, due to the ever-growing amount of different entities and/or product with similar characteristics, the users are not only aiming at analyzing and searching the best entities that match their interests but their goal is also to find those that agree

with the vast of majority of the users who already have an opinion about the considered entity. In other words, they are often searching for authentic, user-generated reviews to orient their search and decisions.

The platform we aim to build will provide an environment for real users and/or enterprises to analyze, search, detect and visualize real world entities within a dynamic stream of massive data and organize the list of relevant results based on knowledge and contextual information.

Acknowledgments. I would like to acknowledge all PhD students, postdoctoral researchers and internships having worked in the team during the past five years.

References

1. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): Mastering the Information Age: Solving Problems with Visual Analytics, Thomas Müntzer (2010)
2. Trujillo, J., Maté, A.: Business Intelligence 2.0: A General Overview. In: Aufaure, M.-A., Zimányi, E. (eds.) eBISS 2011. LNBIP, vol. 96, pp. 98–116. Springer, Heidelberg (2012)
3. Kobsa, A.: Generic user modeling systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 136–154. Springer, Heidelberg (2007)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (2001)
5. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman & Hall/CRC (2009)
6. Aggarwal, C. (ed.): Data Streams. Models and Algorithms. Advances in Database Systems, vol. 31. Springer (2007)
7. Tatbul, N., Cetintemel, U., Zdonik, S.: Staying FIT: Efficient Load Shedding Techniques for Distributed Stream Processing. In: International Conference on Very Large Data Bases (VLDB 2007), Vienna, Austria (2007)
8. Ganter, B., Wille, R.: Formal Concept Analysis. Mathematical Foundations Edition. Springer (1999)
9. Soussi, R., Cuvelier, E., Aufaure, M.A., Louati, A., Lechevallier, Y.: DB2SNA: an All-in-one Tool for Extraction and Aggregation of underlying Social Networks from Relational Databases. In: Ozyer, T., et al. (eds.) The Influence of Technology on Social Network Analysis and Mining, Springer (2012) ISBN 978-3-7091-1345-5
10. Buitelaar, P., Cimiano, P. (ed.): Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Series Information for Frontiers in Artificial Intelligence and Applications. IOS Press (2008)
11. Ben Mustapha, N., Aufaure, M.A., Baazaoui-Zghal, H., Ben Ghezala, H.: Query-driven approach of contextual ontology module learning using web snippets. Journal of Intelligent Information Systems (2013)
12. Tiddi, I., Mustapha, N.B., Vanrompay, Y., Aufaure, M.-A.: Ontology Learning from Open Linked Data and Web Snippets. In: Herrero, P., Panetto, H., Meersman, R., Dillon, T. (eds.) OTM-WS 2012. LNCS, vol. 7567, pp. 434–443. Springer, Heidelberg (2012)
13. Giacometti, A., Marcel, P., Negre, E.: A framework for recommending OLAP queries. In: Proc. DOLAP, Napa Valley, USA, pp. 73–80 (2008)
14. Aufaure, M.-A., Kuchmann-Beauger, N., Marcel, P., Rizzi, S., Vanrompay, Y.: Predicting your next OLAP query based on recent analytical sessions. In: Bellatreche, L., Mohania, M.K. (eds.) DaWaK 2013. LNCS, vol. 8057, pp. 134–145. Springer, Heidelberg (2013)

15. Hu, B., Vanrompay, Y., Aufaure, M.-A.: PQMPMS: A Preference-enabled Querying Mechanism for Personalized Mobile Search. In: Faber, W., Lembo, D. (eds.) RR 2013. LNCS, vol. 7994, pp. 235–240. Springer, Heidelberg (2013)
16. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
17. Cataldi, M., Ballatore, A., Tiddi, I., Aufaure, M.A.: Good Location, Terrible Food: Detecting Feature Sentiment in User-Generated Reviews. International Journal of Social Network Analysis and Mining (2013)
18. Wasserman, S., Faust, K.: Social Network Analysis: Methods and applications. Cambridge University Press (1994)
19. Kuchmann-Beauger, N., Brauer, F., Aufaure, M.A.: QUASL: A Framework for Question Answering and its Application to Business Intelligence. In: Seventh IEEE International Conference on Research Challenges in Information Science (2013)
20. Thollot, R., Kuchmann-Beauger, N., Aufaure, M.-A.: Semantics and Usage Statistics for Multi-Dimensional Query Expansion. In: Lee, S.-g., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (eds.) DASFAA 2012, Part II. LNCS, vol. 7239, pp. 250–260. Springer, Heidelberg (2012)