Big Data: an Information Systems approach

Jacky Akoka, Faten Atigui, Isabelle Comyn-Wattiau, Fayçal Hamdi, Elena Kornyshova, Nadira Lammari, Elisabeth Métais, Cédric du Mouza, Nicolas Prat, Samira Si Saïd-Cherfi

firstName.lastName@cnam.fr

CNAM, Paris, France

Abstract. Several challenges and issues characterize the research on big data including in social media data, mobile data, web data, and network data. While much has been written in terms of technologies and algorithms, much less has been written about methodologies and frameworks that could enable researchers and practitioners to more efficiently tackle the big data challenges and issues they face. The aim of the ISID Group at CNAM CEDRIC is to propose methodologies, approaches, and frameworks in order to cope with these challenges. This paper presents issues of the Big Data research from the Information System point of view and describes different research topics of the ISID Group.

Keywords. Big Data, Information System, Reverse engineering, Process methodology, Security, Quality, Data Warehouse, Social Web.

Several challenges and issues characterize the research on Big Data. Let us mention some of these issues: 1) There is clearly a need to define a big data process methodology. 2) Big data has challenging security and privacy problems requiring new approaches. Let us remind that privacy and security are the most important big data issues including conceptual significance. Big data anonymization and privacy can be combined in an integrated framework. 3) Given the range of big data applications, there is a need to develop methodologies to tackle the data quality issues. It is obvious that bad data quality in the context of big data can lead companies to disastrous situations. Big data quality needs to be conceptualized in a framework. 4) Approaches for data integration issues that arise in many real-life settings are needed. 5) Issues of big data information provenance require new approaches. 6) Since big data is unstructured, traditional

analysis methods are insufficient to analyze huge volume of data. 7) Representation of unstructured data coming from different sources require the use of specific methods. 8) Theoretical foundations for big data especially based on an ontology is missing. 9) An ontology-based approach to big data analytics remains to be developed. It can be used to address the semantic challenges presented by big and unstructured data sets. 10) Framework for big data-driven risk management is yet not available. 11) A conceptual model for big data is not well developed. 12) Big data warehouse conceptual model, based on MDA, has been proposed but not fully tested and validated.

Reverse engineering and Integration of Big Data

Chikofsky and Cross [Chikofsky90] defined reverse engineering to be "analyzing a subject system to identify its current components and their dependencies, and to extract and create system abstractions and design information." Existing reverse engineering methods and tools focus on extracting the structure of a legacy system with the goal to reengineer or to reuse it. In the context of big data, incorporating data from various sources (internal and external) in a data warehouse or in a big data warehouse requires a data model. However the variety of data, its volume, and its velocity create problems difficult to solve. Conceptual modeling of big data appears to be one of the challenges facing researchers. Various conceptual modeling techniques, such as ontologies, semantic, RDF schemas, SPARQL Language, etc., are not well suited for big data conceptualization. We propose a new approach based on reverse engineering and schema integration techniques.

Processes and Methods in Big Data

Situational Method Engineering (SME) offers a wide range of approaches allowing to adapt methods and processes to a given situation: method family configuration [Kornyshova11], method extension [Deneckère01], contextual reconstitution from method components [Ralyté01] etc. Information Systems engineering methods should be revised to fit the context of Big Data and to be able to deal with

scalability, velocity, variety, variability and other Big Data issues. Our approach to manage this problem is to use the SME techniques to adapt different processes and methods in the context of Big Data in the following manner. Methods or processes conceived for the same purpose are analyzed to identify the common and variable components, then a meta-method (or a meta-process) is constructed from the initial ones, finally a set of guidelines is suggested to define a method or to execute a process adapted to a given case. These contextual method definition and process execution are done using decision-based guidance.

Data and Data Interlinking Quality

Several Data and Linked Open Data issues are actually investigated within the ISID team: The identification of quality defects related to data and data interlinking quality. Several recent contributions from literature pointed out the lack of quality [Halpin10, Zaveri15]. The challenge is to investigate several interlinked data sources from several domains and try to qualify the underlying quality defects. Once quality criteria identified, it is necessary to associate to each criterion a set of assessment methods and algorithms. The detection of quality defects raises the problem of correction that is not sufficiently addressed in literature where quality stands more for evaluation than correction. The proposed solutions should be implemented through a prototype or a platform to support Data Interlinking evaluation and improvement. Finally, the problem of scalability should be addressed as Open Data is also Big Data and developed solutions should be scalable.

Data Security & Privacy

Following a security incident, a security analysis is often required. It focuses first on the traces of computer systems (logs) in order to reconstruct what has happened and deduce the attacker's mode of operation. Due to the explosion of connected objects and the proliferation of online and surfing activities carried out on social networks, Internet induces the appearance of numerous and various traces. To cope with this enormous amount of data, security analysts have

equipped themselves with computer tools known as SIEM (Security Information and Event Management). The latter, given the difficulty of automating the complex process of human reasoning, generate confusing results and multitude of false positives and false negatives. For instance, the HuMa project will capitalize on the complex reasoning of security experts in order to introduce guidance in SIEM tools and to permit to security analysts to react on the fly.

Data Warehouse Modeling

We have worked on data warehouse and dimensional modeling for many years. At the era of big data warehousing, we aim to define logical models for each noSQL database family (column-based, graph, key-value, and document) and mapping rules between UML/logical and physical noSQL levels. Until now, the noSQL community does not refer to logical or conceptual models. We claim that adding such models will considerably facilitate the handling of big data warehouses.

Social Web and Recommendations

Micro-blogging platforms such as Twitter, Pinterest, Instagram, Weibo or Tumblr all share this mechanism of selecting interesting people to follow and being followed by other users which is now well established in the Internet culture. But with this success, microblogging platforms began to be very crowded and users started having issues to keep up with all the content available. We investigate specific collaborative filtering methods to recommend items of interest, based on both content and topology of the underlying social network, which can scale up to very large datasets [Constantin16]. In this context we have also investigated edge-partitionning solution to allow intensive graph-computation on very large graphs [Li16]. Finally we have proposed in [PCM16] a distributive collaborative filtering system based on the semantic knowledge of the domain in order to help the user in the e-commerce area. The main issue with this approach concerns cold-start. In [PCM16a] we have used bloom filters and parallelization to deal with scalability.

References

[Chikofsky90] E. Chikofsky and J. Cross, Reverse engineering and design recovery: A taxonomy. IEEE Software, 7(1):13- 17, January 1990

[Constantin16] C. Constantin, R. Dahimene, Q. Grossetti, C. du Mouza: Finding Users of Interest in Micro-blogging Systems. EDBT 2016: 5-16

[Deneckère01] Deneckere R., Approche d'extension de méthodes fondée sur l'utilisation de composants génériques. PhD thesis. University of Paris 1 Panthéon-Sorbonne, 2001.

[Halpin10] Halpin, H., Hayes, P. J., McCusker, J. P., McGuinness, D. L., & Thompson, H. S. When owl: sameas isn't the same: An analysis of identity in linked data. In The Semantic Web–ISWC 2010, 2010

[Kornyshova11] Kornyshova E., Deneckère R., and Rolland C. Method Families Concept: Application to Decision-Making Methods, EMMSAD, London, United Kingdom, 2011.

[Li16] Y. Li, C. Constantin, C. du Mouza: A Block-Based Edge Partitioning for Random Walks Algorithms over Large Social Graphs. WISE (2) 2016

[PCM16] M. Pozo, R. Chiky, E. Metais. "Enhancing collaborative filtering by using implicit relations in data", LNCS Transactions on Computational Collective Intelligence (TCCI), vol. 9655(.), 2016

[PCM16a] M. Pozo, R. Chiky, F. Meziane, E. Metais. An Item/User Representation for Recommender Systems based on Bloom Filters, RCIS, 2016.

[Ralyté01] Ralyté J., Rolland C., An Assembly Process Model for Method Engineering. In: Proceedings of CAISE 2001, Springer, Berlin, 2001.

[Zaveri15] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. Quality assessment for linked data: A survey. Semantic Web, 7(1), 2015.