Big data for Enhanced Learning Analytics: A case for large-scale comparative assessments

Nikolaos Korfiatis

korfiatis@em.uni-frankfurt.de

Big Data Analytics Research Lab, Chair for Database and Information Systems,
Institute for Informatics and Mathematics, Goethe University Frankfurt
Robert-Mayer-Str. 10, 60325, Frankfurt am main, , Germany
http://www.bigdata.uni-frankfurt.de

Abstract

Recent attention on the potentiality of cost-effective infrastructures for capturing and processing large amounts of data, known as *Big Data* has received much attention from researchers and practitioners on the field of analytics. In this paper we discuss on the possible benefits that *Big Data* can bring on TEL by using the case of large scale comparative assessments as an example. Large scale comparative assessments can pose as an intrinsic motivational tool for enhancing the performance of both learners and teachers, as well as becoming a support tool for policy makers. We argue why data from learning processes can be characterized as *Big Data* from the viewpoint of data source heterogeneity (variety) and discuss some architectural issues that can be taken into account on implementing such an infrastructure on the case of comparative assessments.

Keywords: Bigdata, TEL, Learning Analytics, Comparative assessments

1 Introduction

Current advances in the domain of data analytics refer more and more to the case of Big Data as the foremost pillar of any modern analytics application [1]. While the appearance of the term *Big Data* in the literature spans multiple definitions, a definition of *Big Data* that is appropriate in the context of user activity modeling can be the one provided by the McKinsey research report [2] which defines *Big Data* as "*datasets*" which for practical and policy reasons cannot be "*processed, stored and analyzed"* by traditional data management technologies and require adaptation of workflows, platforms and architectures. On the other hand, research on the field of technology enhanced learning (TEL) highlights the importance of dataset-driven research for evaluating the effectiveness of technological interventions on the learning process. As it is also in the case of traditional data driven evaluation and decision making, data analytics for TEL considers the combination of data from various sources centered on the educational process lifecycle.

The view that we advocate on this position paper is the belief that the definition of *Big Data* itself contains higher level semantics, which under particular application domains (in that case TEL) have different applicability for data infrastructures and research information systems. While someone can argue that the order of magnitude (*volume*) of most datasets generated in TEL is something that can be handled by traditional tools (e.g. a typical database size of 4-5 Gigabytes is not considered problematic for a traditional DBMS), the issue arises when the *variety* of sources where data can be integrated (e.g. log files, assessment scores, essays etc.) comes in consideration. In addition, when these insights are targeted to the stakeholders (learners, teachers and policy makers) by encompassing some type of interactivity (e.g. through visualizations) the case of responsiveness on providing the insights related with the computation time that it takes for such an insight to be calculated is a contributing factor. We frame this factor as *velocity*. We discuss on the interplay between those three factors (*volume, variety, velocity*) by adopting the well-known 3V architectural model of Big Data[3] and introducing an application scenario on the case of large scale comparative assessments, inspired by the recent attention of policy makers on comparative rankings such as the PISA report [4].

Our viewpoint is that big data can greatly enhance TEL and contribute insights on better assessing the results of technological interventions on the learning process. To this end this paper is structured as follows. A discussion of big data architectures is highlighted in Section (2). A subsequent framework is introduced on the application of big data in learner settings on Section (3). The paper concludes on Section (4) with discussion and directions for further research.

2 Big data on integrating insights from learning analytics

2.1 Data Source heterogeneity in TEL

A majority of studies on learning analytics consider the individual as a basic unit of analysis. Nonetheless learners' participation from a data analysis point of view has been addressed as a problem of combining heterogeneous data sources [5]. From that perspective learning analytics considers the following three interconnected stages with reference to modeling the learners' participation in educational activities and in particular namely: (a) User knowledge modeling, (b) User behavior modeling and (c) user experience modeling.

In essence the emphasis is given on two major contributors of different types of data that support users of analytics tools on gaining interesting insights about the learners: (a) *Social Interactions data*: providing the interactions the learner has with the educators and fellow students as well as navigation in the learning platform, and (b) *Social or "attention" metadata*: that determine which aspect of the learning process the learner is actively engaged with. Social interaction data can be considered as a facet of data containing standard user navigation properties, similar to the ones that standard Business Intelligence (BI) tools provide in a corporate environment to an analyst. The latter can be seen in cases such as web usage mining, content navigation extraction etc. and are considered as a mature source for analytics applications since are not difficult to extract from standard application logs (e.g. web server logs for web usage mining). On the other hand attention metadata require an interaction with a learning resource as in the context of a learning object repository, where the learner interacts with a learning object in the context of an activity (e.g. an exercise or an interactive tutorial) and are more difficult to extract due to the fact that in order for them to be generated an interaction has to be recorded. Similar to research issues in other domains such as online communication, the case of "shadow" participation of users, labeled as "lurking" [6], makes the extraction of a relatively large enough dataset from social interaction data quite difficult to achieve.

For example consider the case where a learning repository provides the opportunity to learners to evaluate the quality of learning objects. Published studies concerning participation on online repositories [7] have shown that as in any other case of user generated content, participation of learners on the evaluation of learning objects follows the "Pareto" or 80/20 rule [8]. That provides that roughly 20% of the participants are responsible for the 80% of the "attention metadata" generated by this interaction. The above makes the case of studying learners in a highly vertical set (e.g. in the context of one particular educational domain/activity) inefficient since the quantitative characterization of the learners' activity might contain hidden biases which is not uncommon for quantitative studies. Therefore enhancing learning analytics through the use of big data can allow for metrics from different learning repositories to be aggregated in order to allow for accurate and non-biased quantitative studies.

2.2 Variety on the 3V Architectural model

As aforementioned Volume, Variety and Velocity constitute an integral aspect of *Big Data* in relation with the different goals that an end user wants to achieve by utilizing large dataset. The interplay of these three distinctive elements has been adapted on the 3V architectural model. The 3V model was first described by Laney [9] on a Gartner Report and considers the three characteristics as a case of interconnected stages [3].

In traditional storage systems the volume (size) defines the amount of data which the system can manage whereas the velocity (speed) defines the speed with which this data is processed. In other words, in OLAP systems the data size can become very large but the data processing speed is slow whereas in OLTP systems the workload size is much smaller, but the data should be processed much faster. The variety (structure) is of no concern in such systems as the data format is known in advance and described very precisely in a pre-defined schema. However, in the age of Big Data the picture is changing and the emerging data variety is starting to transform the system requirements and question the storage efficiency of existing systems.

A particular requirement is that a system from its foundations should be capable of handling increasing data volume, high or dynamically changing velocity and high variety of data formats like the ones that can be combined in different cases of TEL platforms featuring different nature of educational content (e.g. the use of screencasts). In the multitude of different TEL scenarios, the exact significance of each of the 3Vs can vary

depending on the requirements and the current TEL platforms should be able to deal with any combination of the 3Vs from a storage perspective.

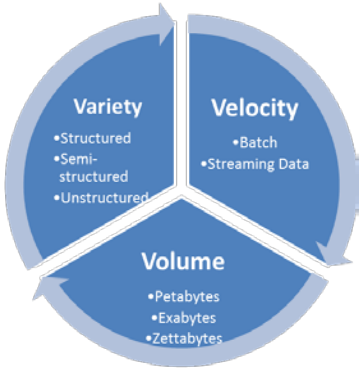

Fig. 1. The 3V architectural model of Big Data (Adopted from [3])

What is interesting with the contribution of the 3Vs as an architectural platform for *Big Data* in TEL is the dynamic changing significance of its elements which depicts in a very abstract way the actual challenge of learning analytics in TEL which depending on the goal and unit of analysis can be different.

3 Utilizing big data on providing comparative assessments in TEL

3.1 Case for support: Collective effort and individual-to-group comparison

A case of support for comparative assessments as a tool to enhance the motivation of learners and teachers in a TEL scenario can be considered the comparison of individuals to the average or the norm. A particular model that has been applied in cases where users massively participate in a common activity (as in the case of online communities) is the collective effort model. The collective effort model has been introduced in social psychology by Karau and Williams [10] and builds on principles from expectancy theory [11]. According to these principles, an individual needs to feel that his/her effort will lead to a collective level of performance in the context of the group. For example in an online community a participant might start posting and participating in the community activities if he/she believes that he will gain a higher status or visibility in the community. **Fig. 2** provides a depiction of the stages of the collective effort model and how it affects individual motivation.

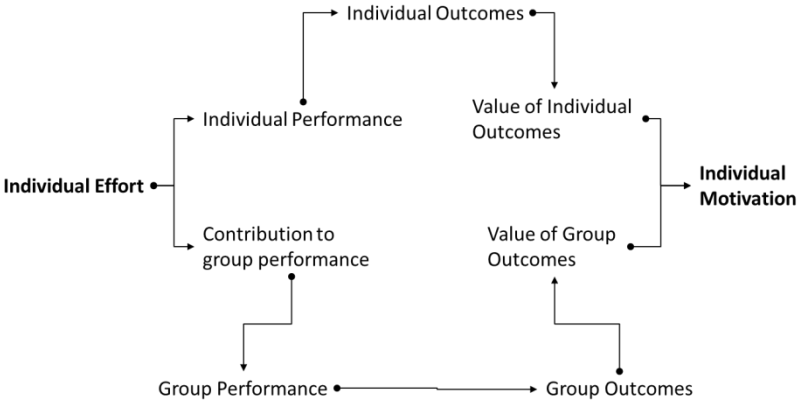

Fig. 2. The collective effort model. Adapted from Karau and Williams [10]

In the middle stage of the process the individual compares his/her effort with the effort of the group. In case he performs lower than the group then an intrinsic effect to contribute more is witnessed. However in the case the performance comparison between the individual and the group outcome is lower, the opposite effect might take place. In that case, it is more likely that the individual effort will be dropped if the comparison of the individual outcome is higher in relation with the group outcome.
An interesting application of the collective effort model has been undertaken by Ling et al. [2005] where the contributions in terms of individual reviews and ratings were used by the MovieLens recommender system [Miller et al., 2003] in order to provide recommendations to other members of the community.
It is our belief that the collective effort model can provide a theoretical case for support for the use of comparative assessments alongside standardized pedagogy models. Comparative assessments can have a bimodal influence on the case of TEL. For example considering the collective effort model for learners, socio-psychological cues provided by such a comparison can influence intrinsically their participation level. For example a learner that has very good performance in class can think: *"….I am the best in my class, so I don't need to devote more time..."* However after making possible to compare his/her performance with a very large number of other learners, in case he is not on the top this can have a motivational factor as: *I might be the best in my class but when compared with other students from other schools I am not so good*
Taking this into account we believe that a set of pedagogical approaches adapted on different learner modalities need to be undertaken in order to study how performance comparison metrics can be federated and how these metrics can encapsulate different levels of granularity for individual-to-individual and individual-to-group comparisons (e.g. school level, region level, country level). Nonetheless this scenario also requires a rethinking on the data integration process from different sources. We highlight this case on the section that follows.

3.2 Integrating data from different learning repositories

As aforementioned in the previous section, a critical stage for encapsulating the intrinsic value of comparative assessments considers the cognitive ability of the individual to evaluate his performance in the context of a common activity with others. However this requires a combination and normalization of data from different environments which might be fragmented due to privacy and policy issues. In that case we believe a second argument for *Big Data* in the case of TEL can be found on helping to overcome the technical and organizational boundaries.
In traditional scenarios where more than one content platform is involved, learning analytics involve the extraction, transform and load (ETL) of data from the learning content repositories to a centralized server, where an analysis is carried out and the output is presented to the stakeholder. We believe that this approach has two major pitfalls:

- *The inefficiency of having to extract the data from their physical location* and move them to a centralized infrastructure where analytics will be performed. This results to inefficiency both from the side of computation as well as from the side of data analysis (metrics in different scales, not easy to replicate the analysis with new data etc.)
- *The organizational and policy implications of having to extract data involving user generated content and activity* which on the one hand are crucial for analysis and comparisons but on the other hand involve open issues in security, privacy and compliance with the legal framework under each provider operates.

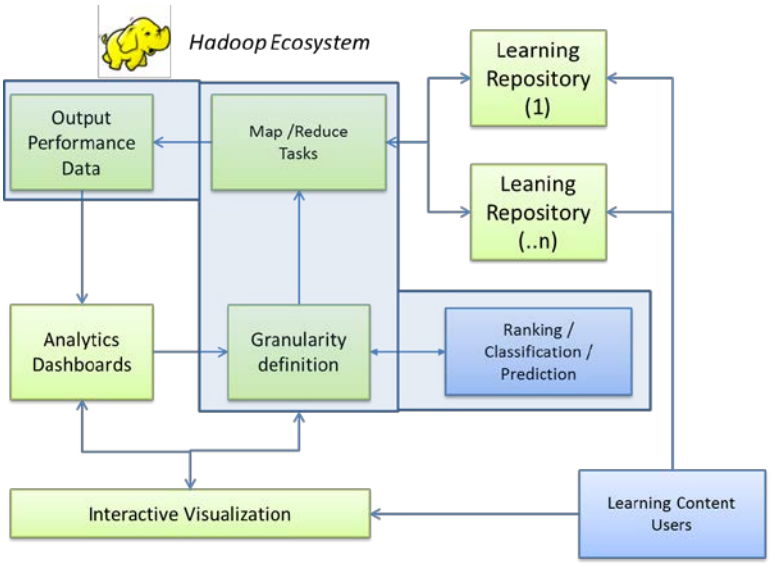

Fig. 3. An example data workflow for providing comparative assessments in TEL with the use of

Fig. 3 provides an example data flow where big data can be utilized for such case. A set of different learning platforms/repositories can be aggregated using the *Big Data* ecosystem in an effective way. The core of such a process considers Apache Hadoop and its ecosystem. This will allow for:

- *Contextualized mapping of the data to be used*, by taking into account the heterogeneity or variety of the learning content residing in different repositories as well as the pedagogical strategies involved from the side of evaluation.
- *Computing and aggregating analytics on the physical point where the data reside* thus making the computation more efficient and allowing the infrastructure to provide interactive data exploration and visualization workflows in various levels of granularity (in day and level of analysis) defined by the user with a pre-anonymized procedure in the analytics extraction process.

In particular the Hadoop ecosystem [12] and its flexibility for data analytics is making a basic implementation of the map/reduce or M/R paradigm introduced by Google [13]. In a very simple implementation an M/R analytics framework consist of two separate and distinct tasks: (a) the Map job, which takes a set of data and converts it into another set, where individual elements are broken down into tuples (key/value pairs) and (b) the Reduce job which takes the output from a map task as input and combines those data tuples mostly into a smaller set of tuples by e.g. presenting aggregates. In a learning object repository integration scenario based on this approach a typical M/R process could consist of:

- *A mapping stage declaring the types of data (interaction data, attention metadata) to be collected from each repository* allowing for contextualized modeling based on the architecture and semantics of the learning platform (something addressed in the context of linked open data).
- *A reducing stage which allows the reduction of the raw data and aggregation of insights in a centralized output* that can be used for presentation or visualization purposes.

The above constitutes a case for support from a technical/infrastructure viewpoint where *Big Data* can advance the state of the art in learning analytics. Our view is that this requires major architectural changes in the support of existing tools and platforms used in TEL with the integration of the new technologies based on the M/R paradigm. This will allow for an efficient way of processing large TEL datasets using existing infrastructures from a cost and organizational point of view [14].

4 Conclusions and Outlook

In this paper we intended to illustrate a case for support for Big Data and the technological ecosystem that accompanies it, with a case study of an educational scenario in learning analytics and in particular the case for providing comparative assessments.

While it is generally accepted that research on the field of learning analytics has advanced our understanding of the effectiveness of using technology tools and platforms for enhancing the educational learning process, it is our inherent belief that *Big Data* can take this understanding to a next level. Future research could also consider cases related with user interaction dynamics [15, 16] when interacting in online platforms, comparative evaluation of learning resources [17] as well as evaluating textual feedback through opinion mining [18].

Nonetheless, building such a competence in the current state of development of big data requires better understanding of the use cases and scenarios where big data can be of use in the study of TEL. While we have highlighted the potentiality of big data in the case of providing comparative assessments in order to enhance participation by learners and performance by teachers, we believe that additional pedagogies and motivational paradigms can be redefined in the same way. In that direction we believe that a closer collaboration between data scientists and experts in pedagogies could produce new results in that direction.

References

1. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. 52, 21–32 (2011).
2. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity. Mckinsey Glob. Inst. 1–137 (2011).
3. Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. (2011).
4. Margaret, W., Ray, A.: PISA Programme for International Student Assessment (PISA) PISA 2000 Technical Report: PISA 2000 Technical Report. OECD Publishing (2003).
5. Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Washington Dc Office of Educ. Technology. Us Dep. Education. 1–57 (2012).
6. Nonnecke, B., Preece, J.: Why lurkers lurk. Proceedings of Americas Conference on Information Systems (AMCIS). pp. 1–10 (2001).
7. Sicilia, M.-A., Ebner, H., Sánchez-Alonso, S., Álvarez, F., Abián, A., García-Barriocanal, E.: Navigating learning resources through linked data: a preliminary report on the re-design of Organic. Edunet. Proc. Linked Learn. 2011, 1st (2011).
8. Cechinel, C., Sicilia, M.-Á., Sánchez-Alonso, S., García-Barriocanal, E.: Evaluating collaborative filtering recommendations inside large learning object repositories. Inf. Process. Manag. 49, 34–50 (2013).
9. Laney, D.: 3D data management: Controlling data volume, velocity and variety. Appl. Deliv. Strat. File. 949, (2001).
10. Karau, S.J., Williams, K.D.: Understanding individual motivation in groups: The collective effort model. Groups Work Theory Res. 113–141 (2001).
11. Christiansen, B.A., Smith, G.T., Roehling, P.V., Goldman, M.S.: Using alcohol expectancies to predict adolescent drinking behavior after one year. J. Consult. Clin. Psychol. 57, 93 (1989).
12. Apache Hadoop: https://hadoop.apache.org/.
13. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. Acm. 51, 107–113 (2008).
14. Leo, S., Anedda, P., Gaggero, M., Zanetti, G.: Using virtual clusters to decouple computation and data management in high throughput analysis applications. Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference on. pp. 411–415. IEEE (2010).
15. Wu, P.F., Korfiatis, N.: You Scratch Someone's Back and We'll Scratch Yours: Collective Reciprocity in Social Q&A Communities. J. Am. Soc. Inf. Sci. Technol. Forthcoming. (2013).
16. Korfiatis, N., Sicilia, M.A.: Social Measures and Flexible Navigation on Online Contact Networks. Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–6. , Imperial College, London UK (2007).

17. Papavlasopoulos, S., Poulos, M., Korfiatis, N., Bokos, G.: A non-linear index to evaluate a journal's scientific impact. Inf. Sci. 180, 2156–2175 (2010).
18. Korfiatis, N., García-Bariocanal, E., Sánchez-Alonso, S.: Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. Electronic. Commerce. Research and. Applications. 11, 205–217 (2012).