Introduction to Big Data and Analytics: Pathways to Maturity "The Original Big Data and Analytics Minitrack"

Stephen H. Kaisler, D.Sc.
SHK & Associates
Laurel, D 20723
Skaisler1@comcast.net
(301) 498-4244

Frank J. Armour, Ph.D.
Kogod School of Business
American University
Washington, DC 20016
fjarmour@gmail.com
(202) 251-3554

J. Alberto Espinosa, Ph.D.
Kogod School of Business
American University
Washington, DC 20016
alberto@american.edu
(202) 885-1958

The Big Data and Analytics Minitrack ("The Original Big Data Minitrack") accepted six papers having to do with all aspects of big data and all types of analytics that can be applied to it. This year the minitrack will present five papers in virtual sessions via Zoom on a variety of topics.

The first paper, by Jeffrey Saltz and Nicholas Hotz, is entitled "Factors that Influence the Selection of a Data Science Process Management Methodology: An Exploratory Study". Selection of a process methodology for managing and coordinating data science project is a critical factor in its success. In their literature survey, they found that 82% of data scientists did not follow a defined process, yet most felt their results could be improved if they used a systematic process methodology. The authors reviewed the six common DS methodologies: Kanban, Scrum, Research-Agile Hybrid, Water-Agile Hybrid, CRISP-DM, and Ad Hoc. Through a literature survey, they identified factors that characterized these methodologies from which they generated nine hypotheses. They used the Technology-Organization-Environment (TOE) framework to organize the hypotheses into three categories and to analyze the data collected from semi-structured interviews across 14 organizations. The organizations varied in size from a 10-person consulting firm to a 250,000-person financial services firm. All but one of their hypotheses was supported by data collected during the interviews. Of the remaining eight hypotheses, all but one had corroboration from at least four of the interviewees. Organizational and environmental factors were particularly relevant and decisive in selecting an approach to project management. The authors conclude that a more detailed drill down into these factors will identify relative importance, cross-factor relationships, and weaknesses, and provide some guidance on new or hybrid methodologies.

Our second paper by Sampsa Suvivuo, entitled "Qualitative Big Data's Challenges and Solutions: An Organizing Review", addresses the challenges in using qualitative big data for analyzing problems. The author notes that 80% or more of organization's data is qualitative or unstructured text, audio, video, and images. Some of the challenges that affect quantitative data are exacerbated when dealing with qualitative data. The author identifies common challenges, but notes they are more difficult to resolve with qualitative data. A key finding is that researchers often use quantitative methods to analyze qualitative data. This approach often requires encoding the qualitative data before analyzing it. Unstructured corpora often have to be organized and formatted into structured corpora before

URI: https://hdl.handle.net/10125/70726
978-0-9981331-4-0

being quantitatively analyzed which imposes extra steps in the analytical process. The author noted the lack of qualitative tools which can limit to analytical results. Identifying and understanding "noise" in qualitative data is a particularly challenging aspect in analyzing such data. Sorting through the mass of data to determine what is relevant and what is not can be very time-consuming and often involves subjective judgments.

This paper identifies a persistent problem in handling unstructured, non-numerical data which has been deferred too long (in our opinion). Renewed emphasis on the development of qualitative tools and their supporting methods is needed to handle the ever-growing amount of qualitative data.

[Note: With the amount of qualitative data increasing as a proportion of big data, there is an urgent need for advanced qualitative tools not dependent on numerical and statistical methods. The co-chairs addressed this problem in their papers [1,2]].

Our third paper by Lucas Baier, Vincent Kellner, Nicholas Kuhl, and Gerhard Satzger, entitled "Switching Scheme: A Novel Approach for Handling Incremental Concept Drift in Real-World Data Sets", The authors address a critical problem in machine learning – concept drift – which occur when the underlying data and principles of a situation change over time. This means that machine learning cannot be a static process, but must dynamically respond to concept drift to update the features of the situation it has learned. The authors propose a mechanism – a switching scheme – which involves retraining and updating of a machine learning model. The authors define concept drift carefully and identify several algorithms for detecting it – STEPD, ADWIN, and HDDDM. They examine methodologies for drift handling as the basis for developing and proposing the switching scheme adaptation strategy. After initially training a model and use it to make predictions. As time passes, they check new predictions and, if drift is detected, the model is retrained and updated. This process can be continuous. It addresses one of the key problems in machine learning and artificial intelligence systems, e.g., the onset of fragility as time passes and the environment in which the system is to be used dynamically changes. Using taxi demand data from New York City, they develop a baseline using existing static models, and then examine the use of adaptation strategies. They conclude that the switching scheme offers significant improvement in prediction results over time by leveraging the strengths of frequent retraining and frequent incremental updating.

This paper provides a new mechanism for handling a persistent problem in AI and ML systems that, heretofore, required extensive manual intervention. This approach demonstrates initial steps in automating the management of the concept drift problem in real world systems.

Our fourth paper, by Manel Souibgui, Faten Abigui, Sadok Ben Yahia, and Samira Si-Said Cherfi, entitled "IRIS-DS: A New Approach for Identifiers and References Discovery in Document Stores", addresses the problem of resolving different forms and names of data in NoSQL databases that refer to the same entity. This problem has long been recognized in ontology mergers. It is exacerbated in NoSQL DBs because there are no schema with information on fields such as data types, structures, and lengths. The authors focused on automatically discovering fields in different document stores that relate to the same entity. Their approach focuses on identifying candidates in different DBS, identifying candidate pairs, and then resolving these pairs using scoring and pruning rules based on syntactic and semantic information. Since over 80% of data is qualitative in nature, an efficient mechanism for resolving entities across different DBs can reduce the manual effort currently employed to merge DBs.

As the authors note, having the join key pairs a priori has been the basis for previous work. However, automatically finding and validating the join key pairs allows other programs to focus on the resolution aspects of merging DBs. This

capability will be needed in the future as more document stores come online and there becomes a critical need to merge them – either temporarily or persistently – in order to provide information to researchers.

Our fifth paper, by Stephen Kaisler, William Money, and Stephen Cohen, entitled "Forensic Analysis of Failing Software Projects: Issues and Challenges", focuses on understanding how to determine why software projects fail. Much has been written about this problem, but it is anecdotal. The authors begin to address the problem of digging deeper into software project failure and the corollary problem of determining how to mitigate the actions and activities that presage or lead to such failures. It has been noted by several authors that software failure is a trillion dollar problem. They define forensic analysis as "the use of scientifically derived and proven methods to preserve, collect, validate, identify, analyze, interpret, and document the evidence derived from digital and other sources for the purpose of facilitating or furthering the reconstruction of events leading to the (impending) failure of a (software) project". Jocularly, we think of this as "Project Autopsy". Through a literature survey as well as the experience of one of the authors (Cohen), the authors have identified numerous challenges and posed issues that need to be addressed in forensic analysis and project recovery. The authors believe this is one of the few attempts to understand and develop models of the causes of software project failure in order to develop corresponding models for failing or failed software project recovery. The authors have organized the challenges into categories: cost, schedule, technology, and functionality. Based on this initial analysis, the authors propose an initial model of factors that can be assessed in examining a perceived failing project and determining the likelihood of failure. This model will be revised and evaluated in forthcoming research. Alternatively, if a project has been perceived to have failed, the authors believe that the model can be walked back to try and discover where the potential for failure first became apparent.

Our sixth paper, which will not be presented in the virtual session, is by Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen, entitled "Automatically Mapping Ad Targeting Criteria between Online Ad Platforms". This paper focuses on how map demographic criteria based on customers perceived interests to organizations web pages. Such targeted mapping can potentially increase the response of customers to products and services of interest to them and reduce the amount of email and other information sent to them. The savings can accrue at both ends. Customers won't have to wade through emails or see ads that they are not interested in and organizations can more effectively utilize resources, coupled with other applications, to focus on customers with a strong likelihood of reviewing and purchasing a product or service. The authors compare two algorithmic approaches – Word2Vec and WordNet – as a means of improving targeting criteria. Each method, upon evaluation, had strengths and weaknesses. Word2Vec yields a rough approximation for using criteria from one platform applied to another platform (for example, from Google Ads to Facebook Ads), while WordNet is more useful when manual review of the criteria is not feasible. However, at this stage, they also conclude that human review and judgment is still needed in this area.

Upon reviewing the five papers that were selected by the co-chairs and the other reviewers – both external as well as chosen from among the authors of submitted papers to the minitrack – we note the interest in qualitative methods although some evaluation was performed using quantitative methods. It is too early to say that we are on the cusp of increasing researching in qualitatively methods for qualitative data, but we encourage this area of research as critically needed for future big data analysis.

References

[1] S. Kaisler, F. Armour, A. Espinosa, and W. Money. 2014. “Advanced Analytics: Issues and Challenges”, *47th Hawaii International Conference on System Sciences*, Hilton Waikoloa, Big Island, HI

[2] S. Kaisler, F. Armour, A. Espinosa, and W. Money. 2014. “Advanced Analytics: Issues and Challenges”, *Encyclopedia of Science and Technology, 3rd Edition*, IGI Global