Introduction to the Minitrack on Big Data and Analytics: Pathways to Maturity

Stephen H. Kaisler, D.Sc
SHK & Associates
Laurel, MD 20723
Skaisler1@comcast.net

Frank Armour, Ph.D.
Kogod School of Business
American University
Washington, DC
farmour@american.edu

Alberto Espinosa, Ph.D.
Kogod School of business
American University
Washington, DC
alberto@american.edu

Introduction to Minitrack

This minitrack is focused on exploring theory, techniques, applications, and understanding of the maturing field of Big Data and Analytics. We have selected papers that demonstrate innovative approaches to analytics and introduce new concepts.

The first session focuses on business applications of big data analytics. The first paper, by Raphael Grytz and Artus Krohn-Grimberghe is *Business Intelligence & Analytics Cost Accounting: A Survey on the Perceptions of Stakeholders*. This paper focuses on an important aspect of business intelligence (BI) – determining how much it costs to provide this service within an organization. The authors survey three types of stakeholders regarding the artifacts that characterize BI. Users seem to value cost accounting more than developers while managers assessed justification as most important. Since cost accounting was classified as a management tool, the authors suggest that the suboptimal implementation could lead to inaccurate results. The authors suggest that concentrating on improving manager's perceptions will improve the chances of selling CA in an organization. This paper informs the reader of some of the key aspects needed to understand how measure the cost of BI. As a result, the reader may realize that BI is an information asset for management rather than just a consumer of funds.

The second paper by Axel Soto, Cynthia Ryan, Fernando Peña Silva, Tapajyoti Das, Jacek Wolkowicz, Evangelos Milios and Stephen Brooks discuss *Data Quality Challenges in Twitter Content Analysis for Informing Policy Making in Health Care*. As more opinions are transmitted through social media, this medium becomes a means for understanding the sentiments associated with health care. However, effective analysis of the data is hindered by noise and reliability of the data. The authors address these challenges and provide several simple, but effective solutions for mitigating these challenges. They enhanced precision of the retrieved data by continuous monitoring of the Twitter stream and providing keywords in the query to support recall. Enhanced understanding of users was supported by applying machine learning techniques to support user profiling. The authors note a major open challenge is applying natural language processing techniques to the short bursts of data contained in tweets. Finally, they identified some privacy and ethics challenges resulting from analyzing publicly available data which require more study.

The third paper by Yichuan Wang entitled *Exploring Configurations for Business Value from Big Data Analytics in Health Care* addresses how to derive business value from Big Data Analytics (BDA) in the Health Care industry. The paper describes a model for assess the successful use of BDA to assess different configurations of organizational resources and capabilities to achieve high quality care in hospital settings. Different configurations were assessed using the Fuzzy Set Qualitative Comparative Approach. This study focused on the interdependencies among the capabilities and resources. Their analysis identified five configurations that lead to low average access readmission rates in hospitals. The primary measures used were consistency, which measured the degree to which a necessary or sufficient relation between a condition and an outcome was met, and coverage, which measured the empirical relevance of each variable's contribution to the outcome.

The second session features four papers on innovative uses of technology applicable to big data problems. The first paper by Kunpeng Zhang, Shaokun Fan, and Harry Jiannan Wang entitled *An Efficient Recommender System Using Locality Sensitive Hashing* (LSH) showed how to improve recommendations by preserving similarities of data while reducing the data dimensions. Recommender

URI: http://hdl.handle.net/10125/49981
ISBN: 978-0-9981331-1-9

systems have become a critical element in online business retail operations. Their objective is to identify a set of N items from a larger set of similar items related to previous purchases that might entice the user to purchase more goods. Item-based collaborative filtering (CF), a commonly used technique, faces two challenges when N grows very large: time complexity and space complexity. Reducing the dimensions of the features used for search and filtering reduces these complexities. By hashing the features of an item, comparison can be made on a single number as opposed to a complex feature set. The authors developed LSH that is more efficient than a standard hashing algorithm, particularly when applied to real-valued data.

The second paper by Thomas Forss is entitled *Feature Enrichment through Multi-gram Models*. This paper introduces multi-gram cosine similarity classification models to facilitate text enrichment. Thomas tested this new model on a large set of hateful and violent data drawn from the Web. Many text classification methods use feature extraction to reduce the dimensionality of the feature set. Significant reduction can eliminate the need for feature selection. Many text classification methods fare poorly on imbalanced data sets where the data values are highly skewed. The multi-gram approach is reported to perform very well on imbalanced data sets.

The third paper by Shing Doong is entitled *Counting Human Flow with Deep Neural Network. Flow* counting has many applications in human space management. Using the channel state information (CSI) available through the IEEE 802.1 1n standard, the author was able to count the people in a given area. Using deep neural networks, Doong classified the CSI data, determined and analyzed patterns of fluctuation in this data, and used the results to predict flow size. The author describes a step-by-step process for processing, cleansing and analyzing the data that is somewhat agnostic the raw data. This paper requires some critical understanding of neural network technology, but the reader will be rewarded with a powerful technique that should be applicable across many domains.

The fourth paper by Stephen Kaisler, William Money and Stephen Cohen is entitled *Smart Objects: An Active Big Data Approach*. This paper proposes a new software paradigm that incorporates key capabilities of self-definition, self-reflection, and self-modification which extend an object-oriented framework. The key idea is that Smart Objects have intelligence that enables them to respond autonomously to adapt to their environment which consists many sources of stimuli and data as well as other smart objects. Underlying this approach is the concept of computational awareness that informs and initiates internal actions autonomously to internal and external stimuli. This paper addresses a set of issues and challenges that the authors believe must be solved in order to implement this paradigm. Within the domain of (near) real-time streaming data in a Big Data application, the authors believe that Smart Objects can yield significant improvements in curating, processing, analyzing and responding to data.

The minitrack co-chairs want to thank the many reviewers who assisted in reviewing the papers, providing valuable feedback to the authors, and allowed us to select these innovative and informative papers for HICSS-51.