Behaviour Informatics: Capturing Value Creation in the Era of Big Data

Chi-Hung Chi

Digital Productivity Flagship, CSIRO, Australia
chihungchi@gmail.com

Abstract. Under the era of Big Data, people have been exploring ways of realizing value from data that are at their fingertips. However, it is found that while collecting data is not difficult, value creation is often a big challenge. This makes the approach of "collecting data first before knowing what to do with them" questionable. In this presentation, we discuss the current challenges of big data analytics and suggest how behaviour analytics on trajectory data can help to realize value creation from Big Data.

1 Background and Challenges

As we move to the fourth paradigm of computing – data intensive scientific discovery, numerous research efforts have been spent in building huge big data repositories. Together with data mining and machine learning research, it is hoped that better and more intelligent decisions can be made in real time.

This movement is accelerated by the advance in at least three areas. The first one is social network, where people share their views and opinions in public. The second one is cloud computing, which is an on-demand infrastructure that facilitates sharing of data, collaboration among multiple parties, and support for on-demand computational and storage infrastructure services at low cost. The third one is the internet-of-things. With the maturity of sensor technologies, trajectory movement of entities (including human and things) can now be monitored in real time at low cost. However, gaining access to big data is only the starting point. There are still open issues that need to be addressed in the value creation process when dealing with big data.

One result of the big data mega trend is the building of huge data repositories around the world. In Australia, the government has been pushing for sharing bureau data through spatial information platforms. It is true that data are collected and can be made available to users, but how to make sense out of these data practically and economically is still a mystery to be explored. Without value creation, the high maintenance cost of these repositories cannot be justified, and the motivation for data providers to update their data inside will also disappear.

In the past few years, sensors and sensing techniques have been advancing rapidly for real time data collection with good enough accuracy. Cost of deploying these technologies is also becoming low enough to make real-time data tracking of human, animals, and even insects (e.g. honey bees) possible. However, without efficient and

effective ways to integrate and transform these trajectory data and their context information into manageable knowledge, these data are actually burdens instead of potentials to their owners.

It is true that there have been numerous research efforts in data mining and machine learning. However, most of them are focused on theoretical algorithmic study, and much less emphasis is put in the incorporation of semantic domain knowledge (in particular, the semantic definition of interdependence among various data sources) into the data mining and pattern discovery processes, and in the use of the behaviour interior dimensions such as loyalty and purchase power of customers to support self service analytics.

Related to the analytics platform, internet-of-things, service and cloud computing techniques are quite mature, and lots of machine learning algorithms are also widely available in both commercial (e.g. MatLib) and open source ("Project R") packages. However, how to put them together in a single service platform and how to compose them together automatically (this is called the vertical service composition) to provide "intelligence-as-a-service" for a given domain are still open for exploration.

2 Real Time Trajectory Data and Its Challenges in Value Creation

In the era of big data, one new important data source for analytics and value creation is the real-time behaviour trajectory data streams of entities (e.g. human) as well as their context dynamics (e.g. environmental such as air quality) that are captured through internet-of-things and sensors (in particular body sensors such as those from Android wears and location position sensors). Its value creation process is both complex and challenging because these data are in general heterogeneous and interdependent on each other. Furthermore, the potential number of data sources, each describing one measurement view of the behaviour dynamics of an entity/event, is in theory, infinite.

Traditional data mining and machine learning approaches from computer science often try to explore co-occurrence patterns and inter-relationship among trajectory data. However, this is usually done without making full use of the interdependence defined by their implicit semantic meaning and domain knowledge. Heterogeneity of data adds another level of complication because quantification measures such as distance are not uniformly and consistently defined across different data types. On the other hand, although domain experts have full knowledge on the semantics of data, they are often not as knowledgeable as computer scientists when dealing with the real time computation on trajectory data streams. This result in the first challenge, how to use data mining / machine learning techniques and domain knowledge together to effectively define and discover the inter-relationships among different trajectory data sources and to perform effective behaviour analysis.

As trajectory-driven behaviour analytics is gaining its recognition in different business and industry sectors, the expectation of decision makers also goes beyond what traditional analytics that mainly focus on statistical summaries and associa-

tion/patterns discovery of transactional/measurable behaviour exterior dimensions often provide. Ultimately, what decision makers want is the deep insight about the behaviour interior knowledge dimensions of entities, by incorporating domain knowledge into the knowledge discovery processes. As an example, the owner of an online shop wants to know not only the "bestselling products of the week", but also the "loyalty", "purchase power", "experience", and "satisfaction" of customers. This results in the second challenge, how to quantify behaviour interior dimensions from exterior transactional (or physically measured) trajectory data and to discover their inter-relationships and relative importance for effective and efficient behaviour analysis.

3 Research Topics in Behaviour Analytics

To achieve this goal, the following is a list of sample research topics for behaviour analytics:

- Effective and efficient deployment of high resolution location tracking network (using Blue-Tooth LE, WiFi-RFIDs, UWB, and Electromagnetic Field) for entities in both indoor and outdoor environment. This forms the basis for behaviour trajectory data tracking and capturing.

- Semantic enrichment of behaviour trajectory data of entities through aggregation of raw trajectory data with their contextual data dynamics, followed by domain knowledge-driven transformation to form behaviour interior dimensions knowledge. This is the data aggregation, integration, and transformation aspects of behaviour analytics; it incorporates domain knowledge into the behaviour trajectory data to create behaviour interior dimensions knowledge as well as to define the interdependence relationship among them.

- Discovery of interdependence relationship among trajectory-driven behaviour data (exterior) and knowledge streams (interior) using data mining techniques. This addresses the interdependence relationships of trajectory data and knowledge streams from the run-time dynamics aspect.

- Coupling interdependence relationships of behaviour trajectory data and knowledge streams into data mining and pattern discovery processes for deep behaviour understanding and prediction. This gives a much better understanding on why things occur; it also gives potentials for future behaviour prediction.

- Design and implementation of a behaviour analytics service system that serves as a publishing, management and operation platform for: (i) software services, (ii) raw trajectory data services, (iii) semantically annotated behaviour trajectory data services (both individuals and collective), (iv) behaviour knowledge services (both individuals and collective), and (v) infrastructure services. Tools to facilitate composition and orchestration of all these services with QoS assurance using public

cloud infrastructure such as Amazon EC2 should be developed. Also, automatic matching of behaviour trajectory data / knowledge services with machine learning / data mining algorithms based on their features should also be supported on this platform.