Reliability Aware Data Fusion

Sameep Mehta
IBM Research India
New Delhi, India
sameepmehta@in.ibm.com

L. Venkata Subramaniam
IBM Research India
New Delhi, India
lvsubram@in.ibm.com

1. OVERVIEW

Due to ubiquitous sensors (GPS, Accelerometer), easy of use apps (Facebook, Twitter etc), presence of audio & video recording devices and higher internet connectivity, the key characteristics of raw data is changing. This new data can be characterized by 4Vs Volume, Velocity, Variety and Veracity. Moreover, due to popular trend of crowd sourcing or citizen sensors, it is reasonable to assume that people will provide multiple evidence of same event using different data types. For example during a Football match, some people will Tweet about Goals, Penalties etc while others will take a picture and upload it. Although the underlying modalities are different (text and image), the data describes the same event. Such multi modal evidences should be used to strengthen the belief in underlying physical event. Finally, each of the data point will have inherent uncertainty. The uncertainty can arise from inconsistent, incomplete, and ambiguous data as well as the trust worthiness of the user. Similarly, some sources are more reliable than others which will also play a part in overall reliability. The volume, velocity and variety are measurable/observable, however, there is no measure of truthfulness.

Traditionally, CS research has focused on Volume and Velocity. However, multimodal data fusion and reliability are less explored. Through this tutorial will wish to draw the attention of researchers towards these dimensions by presenting some real life use cases, highlighting the key technical challenges, existing techniques and need for new .

2. TOPICS

We intend to over the following topics during the tutorial

- Data Characteristics with 4V dimensions and use cases from Public Safety Domain.
- Key Technical Challenges (non exhaustive list)
  - Entity Resolution
  - Data Cleaning
  - Performance
  - Indexing and Storage
  - Updating of Data
  - Use Case: Generating Single View of Entity
- Data Fusion Methods
  - Probabilistic Data Fusion using Bayes Theorem,
  - Information Measures like Entropy, Mutual Information, Fisher Information
  - Interval Calculus, Fuzzy Logic and Evidential Reasoning
  - Kalman Filters & variants, Nearest Neighbor Filters and Probabilistic Data Association Filter
- Reliability
  - Bayesian Methods
  - Dempster Shafer Theory
  - Transferable Belief Theory
- Recent Work in Data/Information Fusion
- Overview of Public Safety using Crowd Sensors Initiatives (National Technical Challenge by IRL)

3. TARGET AUDIENCE

This tutorial is designed for students and researchers in Computer Science. Elementary knowledge of text mining is assumed. This topic is expected to be of wide interest given its overlap with data mining, text mining, NLP, Streaming Data and BigData. We plan to give a 3 hour tutorial.

4. SPEAKERS

L Venkata Subramaniam manages the information processing and analytics group at IBM Research India. He received his PhD from IIT Delhi in 1999. His research focuses on unstructured information management, statistical natural language processing, noisy text analytics, text and data mining, information theory, speech and image processing. He often teaches and guides student thesis at IIT Delhi on these topics. He co founded the AND (Analytics for Noisy Unstructured Text Data) workshop series and also co-chaired the first four workshops, 2007-2010. He was guest co-editor of two special issues on Noisy Text Analytics in the International Journal of Document Analysis and

The 18th International Conference on Management of Data (COMAD), 14th-16th Dec, 2012 at Pune, India.

Recognition in 2007 and 2009. He can be reached at lvsubram@in.ibm.com.

Sameep Mehta is researcher in Information Management Group at IBM Research India. He received his MS and Ph.D from The Ohio State University, USA in 2006. He also holds an Adjunct Faculty position at International Institute of Information Technology, New Delhi. Sameep regularly advises MS and PhD students at University of Delhi and IIT Delhi. He regularly delivers Tutorials at COMAD (2009, 2010 and 2011). His current research interest includes Data Mining, Business Analytics, Service Science, Text Mining, and Workforce Optimization. He can be reached at sameepmehta@in.ibm.com.