Towards Reliability-Aware Entity Analytics and Integration for Noisy Text at Scale

Sameep Mehta, L. Venkata Subramaniam
IBM Research India
sameepmehta,lvsubram@in.ibm.com

1 Outline

Due to easy to use apps (Facebook, Twitter, etc.), higher Internet connectivity and always on facility allowed by smart phones, the key characteristics of raw data are changing. This new data can be characterized by 4V's - Volume, Velocity, Variety and Veracity. For example during a Football match, some people will Tweet about goals, penalties, etc., while others may write longer blogs and further there will be match reports filed in trusted online news media after the match. Although the sources may be varied, the data describes and provides multiple evidences for the same event. Such multiple evidences should be used to strengthen the belief in the underlying physical event as the individual data points may have inherent uncertainty. The uncertainty can arise from inconsistent, incomplete and ambiguous reports. The uncertainty is also because the trust levels of the different sources vary and affect the overall reliability. We will summarize various efforts to perform reliability aware entity integration.

The other problem in text analysis in such setting is posed by presence of noise in the text. Since the text is produced in several informal settings such as email, blogs, tweet, SMS, chat and is inherently noisy and has several veracity issues. For example, missing punctuation and the use of non-standard words can often hinder standard natural language processing techniques such as part-of-speech tagging and parsing. Further downstream applications such as entity extraction, entity resolution and entity completion have to explicitly handle noise in order to return useful results. Often, depending on the application, noise can be modeled and it may be possible to develop specific strategies to immunize the system from the effects of noise and improve performance. Also the aspect of reliability is key as a lot of this data is ambiguous, incomplete, conflicting, untrustworthy and deceptive.

The key goals of this tutorial are:

Tutorials, NAACL-HLT 2013, pages 7–9,
Atlanta, Georgia, June 9 2013.

1. Draw the attention of researchers towards methods for doing entity analytics and integration on data with 4V characteristics.
2. Differentiate between noise and uncertainty in such data.
3. Provide an in-depth discussion on handling noise in NLP based methods.
4. Finally, handling uncertainty through information fusion and integration.

This tutorial builds on two earlier tutorials: NAACL 2010 tutorial on Noisy Text and COMAD 2012 tutorial on Reliability Aware Data Fusion. In parallel the authors are also hosting a workshop on related topic "Reliability Aware Data Fusion" at SIAM Data Mining Conference, 2013.

2 Outline

2.1 Data with 4V characteristics

- Define Volume, Velocity, Variety and Veracity and metrics to quantify them
- Information extraction on data with 4V characteristics

2.2 Key technical challenges posed by the 4V dimensions and linguistics techniques to address them

- Analyzing streaming text
- Large scale distributed algorithms for NLP
- Integrating structured and unstructured data
- Noisy text analytics
- Reliability
- Use case: Generating single view of entity from social data

2.3 Computing Reliability and Trust

- Computing source reliability
- Identifying Trust Worthy Messages
- Data fusion to improve reliability: Probabilistic data fusion, information measures, evidential reasoning
- Use case: Event detection using social data, news and online sources

3 Speaker Bios

Sameep Mehta[1] is researcher in Information Management Group at IBM Research India. He received his M.S. and Ph.D. from The Ohio State University, USA in 2006. He also holds an Adjunct Faculty position at the 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 interests include Data Mining, Business Analytics, Service Science, Text Mining, and Workforce Optimization.

L. Venkata Subramaniam[2] manages the information management analytics and solutions 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. His tutorial titled Noisy Text Analytics was the second largest at NAACL-HLT 2010. 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 Recognition in 2007 and 2009.

[1] http://in.linkedin.com/in/sameepmehta
[2] https://sites.google.com/site/lvs004/