Towards a Social Media Analytics Platform: Event Detection and User Profiling for Twitter

— A Tutorial at WWW 2014 —

Manish Gupta
Microsoft, India
gmanish@microsoft.com

Rui Li
Yahoo! Inc.
ruililab@yahoo-inc.com

Kevin Chen-Chuan Chang
Univ. of Illinois at
Urbana-Champaign
kcchang@illinois.edu

1. Abstract

Microblog data differs significantly from the traditional text data with respect to a variety of dimensions. Microblog data contains short documents, SMS kind of language, and is full of code mixing. Though a lot of it is mere social babble, it also contains fresh news coming from human sensors at a humungous rate. Given such interesting characteristics, the world wide web community has witnessed a large number of research tasks for microblogging platforms recently. Event detection on Twitter is one of the most popular such tasks with a large number of applications. The proposed tutorial on social analytics for Twitter will contain three parts. In the first part, we will discuss research efforts towards detection of events from Twitter using both the tweet content as well as other external sources. We will also discuss various applications for which event detection mechanisms have been put to use. Merely detecting events is not enough. Applications require that the detector must be able to provide a good description of the event as well. In the second part, we will focus on describing events using the best phrase, event type, event timespan, and credibility. In the third part, we will discuss user profiling for Twitter with a special focus on user location prediction. We will conclude with a summary and thoughts on future directions.

2. Presenters

- **Manish Gupta** is an applied researcher at the Bing team in Microsoft India R&D Private Limited at Hyderabad, India. He is also a visiting faculty at International Institute of Information Technology, Hyderabad. He received his Masters in Computer Science from IIT Bombay in 2007 and his Ph.D. from the University of Illinois at Urbana-Champaign in 2013. He worked for Yahoo! Bangalore for two years. His research interests are in the areas of web mining, data mining and information retrieval. He has published more than 25 research papers in referred journals and conferences, including ICDE, KDD, PKDD, SDM, WWW conferences.

- **Rui Li** is a scientist at Yahoo! lab. Before joining Yahoo! lab, he obtained his PhD in the Data and Information Systems (DAIS) Lab at UIUC in 2013. He was advised by Professor Kevin Chen-Chuan Chang. He received his Bacholer's degree in Computer Science ACM Honor Class from Shanghai Jiaotong University, China in 2007.

- **Kevin Chen-Chuan Chang** is an Associate Professor in Computer Science, University of Illinois at Urbana-Champaign. He received a BS from National Taiwan University and PhD from Stanford University, in Electrical Engineering. His research addresses large scale information access, focusing on "entity-centric" information search, integration and mining over Web and social networks. He received a VLDB Best Papers Selection in 2000, an NSF CAREER Award in 2002, an NCSA Faculty Fellow Award in 2003, IBM Faculty Awards in 2004 and 2005, Academy for Entrepreneurial Leadership Faculty Fellow Award in 2008, and the Incomplete List of Excellent Teachers at University of Illinois in 2001, 2004, 2005, 2006, 2010, and 2011. He loves to bring research results to the real world and, with his students, co-founded Cazoodle, a startup from the University of Illinois, for deepening vertical "data-aware" search over the web.

3. Duration and Sessions

Proposed duration of the tutorial is 3 hours. The entire tutorial will be divided into three sessions: Event Detection, Event Description and User Profiling.

4. Topic and Description

Here is a brief outline of the tutorial.

- Event Detection for Microblogging Platforms
  - Event Detection using Tweet Content: Bursty keywords, Graph community analysis, Locality sensitive hashing, Conditional random fields, Hashtags, Tag correlations, Segments (or semantic phrases) [1, 4, 13, 18]
  - Event Detection using Other External Sources: News, Knowledge Bases [11]
  - Applications of Event Detection: Forest-fires, Sporting events, Local festivals, Drug related adverse events, Traffic events, Epidemics, Earthquakes, Emerging controversial events [7, 12, 20, 23]

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577264.

- Event Description for Microblogging Platforms
  - Finding Best Phrase to Summarize an Event [24]
  - Finding Event Types [21]
  - Finding Event Timespans [19]
  - Finding Event Credibility [2, 9]
- User Profiling for Microblogging Platforms
  - Content Based Profiling: General Content Based Attribute Profiling (Gender, Interests), Location Profiling [3, 5, 8, 10, 17]
  - Network Based Profiling: General Network Based Profiling (Interests, University), Location Profiling [6]
  - Hybrid Approaches [14, 16, 22]
  - Co-Profiling Attributes and Relationships [15]
- Summary and Future Research Directions

5. Audience

Researchers in the field of analysis of microblogging platforms will benefit the most as this will give them an exhaustive overview of the research in the direction of event detection for microblogging platforms. We believe that the tutorial will give the newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more. Folks from the industry will benefit from the discussions on a large number of applications where such mechanisms have already been applied.

After the tutorial, the audience will be able to appreciate and understand the following: (1) How text streams can be processed to identify interesting topics (or events); (2) Challenges in displaying those topics (or events) to users; (3) Challenges in predicting location of events; and (4) Challenges in user profiling for microblogging platform users.

6. Relevance to WWW 2014 Areas

The proposed tutorial is most related to the "Web mining" and the "Social networks and graph analysis" areas of WWW 2014. Besides analysis of other forms of data, both the areas deal with analysis of the Twitter network and data in the form of tweets.

7. References

[1] F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum. EnBlogue: Emergent Topic Detection in Web 2.0 Streams. In *Proc. of the 2011 ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD)*, pages 1271–1274, 2011.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proc. of the* 20^{th} *Intl. Conf. on World Wide Web (WWW)*, pages 675–684, 2011.

[3] Z. Cheng, J. Caverlee, and K. Lee. You are Where you Tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proc. of the* 19^{th} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 759–768, 2010.

[4] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang. Discover Breaking Events with Popular Hashtags in Twitter. In *Proc. of the* 21^{st} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1794–1798, 2012.

[5] N. Dalvi, R. Kumar, and B. Pang. Object Matching in Tweets with Spatial Models. In *Proc. of the* 5^{th} *ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 43–52, 2012.

[6] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo. Inferring the Location of Twitter Messages based on User Relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[7] B. De Longueville, R. S. Smith, and G. Luraschi. OMG, from here, I can see the Flames!: A Use-case of Mining Location Based Social Networks to acquire Spatio-temporal Data on Forest Fires. In *Proc. of the 2009 Intl. Workshop on Location Based Social Networks*, pages 73–80, 2009.

[8] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1277–1287, 2010.

[9] M. Gupta, P. Zhao, and J. Han. Evaluating Event Credibility on Twitter. In *Proc. of the 2012 SIAM Intl. Conf. on Data Mining (SDM)*, pages 153–164, 2012.

[10] B. Han, P. Cook, and T. Baldwin. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. In *Proc. of the* 23^{rd} *Intl. Conf. on Computational Linguistics (COLING)*, pages 1045–1062, 2012.

[11] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan. STED: Semi-supervised Targeted-interest Event Detection in Twitter. In *Proc. of the* 19^{th} *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 1466–1469, 2013.

[12] R. Lee and K. Sumiya. Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection. In *Proc. of the 2nd ACM SIGSPATIAL Intl. Workshop on Location Based Social Networks*, pages 1–10, 2010.

[13] C. Li, A. Sun, and A. Datta. Twevent: Segment-based Event Detection from Tweets. In *Proc. of the* 21^{st} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 155–164, 2012.

[14] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *Proc. of the 2012 IEEE 28th Intl. Conf. on Data Engineering (ICDE)*, pages 1273–1276, 2012.

[15] R. Li, C. Wang, and K. Chang. User Profiling in Ego Network: An Attribute and Relationship Type Co-profiling Approach. In *Proc. of the* 23^{rd} *Intl. Conf. on World Wide Web (WWW)*, pages 675–684, 2011.

[16] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In *Proc. of the* 18^{th} *ACM Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 1023–1031, 2012.

[17] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson. The Where in the Tweet. In *Proc. of the* 20^{th} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 2473–2476, 2011.

[18] M. Mathioudakis and N. Koudas. Twittermonitor: Trend Detection over the Twitter Stream. In *Proc. of the 2010 ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD)*, pages 1155–1158, 2010.

[19] D. Metzler, C. Cai, and E. Hovy. Structured Event Retrieval over Microblog Archives. In *Proc. of the 2012 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 646–655, 2012.

[20] A.-M. Popescu and M. Pennacchiotti. Detecting Controversial Events from Twitter. In *Proc. of the* 19^{th} *ACM Intl. Conf. on Information and Knowledge Management (CIKM)*, pages 1873–1876, 2010.

[21] A. Ritter, O. Etzioni, S. Clark, et al. Open Domain Event Extraction from Twitter. In *Proc. of the* 18^{th} *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 1104–1112, 2012.

[22] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your Friends and Following them to Where you are. In *Proc. of the* 5^{th} *ACM Intl. Conf. on Web Search and Data Mining (WSDM)*, pages 723–732, 2012.

[23] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. of the* 19^{th} *Intl. Conf. on World Wide Web (WWW)*, pages 851–860, 2010.

[24] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 685–688, 2010.