Microblogs Data Management and Analysis

Amr Magdy Mohamed F. Mokbel

Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN, USA
{amr,mokbel}@cs.umn.edu

***Abstract*—Microblogs data, e.g., tweets, reviews, news comments, and social media comments, has gained considerable attention in recent years due to its popularity and rich contents. Nowadays, microblogs applications span a wide spectrum of interests, including detecting and analyzing events, user analysis for geo-targeted ads and political elections, and critical applications like discovering health issues and rescue services. Consequently, major research efforts are spent to analyze and manage microblogs data to support different applications. In this tutorial, we give a 1.5 hours overview about microblogs data analysis, management, and systems. The tutorial gives a comprehensive review for research efforts that are trying to analyze microblogs contents to build on them new functionality and use cases. In addition, the tutorial reviews existing research that propose core data management components to support microblogs queries at scale. Finally, the tutorial reviews system-level issues and on-going work on supporting microblogs data through the rising big data systems. Through its different parts, the tutorial highlights the challenges and opportunities in microblogs data research.**

I. Introduction

Microblogs data, e.g., tweets, reviews, news comments, and social media comments, has become very popular in recent years. Everyday, over billion users post more than four billions microblogs [10], [43] on Facebook and Twitter. Such tremendous amounts of user-generated data have rich contents, e.g., news, updates on on-going events, reviews, and discussions in politics, products, and many others. The richness of microblogs data has motivated researchers and developers worldwide to take advantage of microblogs to support a wide variety of practical applications, including social media analysis [44], discovering health-related issues [33], real-time news delivery [4], rescue services [9], and geo-targeted advertising [31]. The distinguished nature of microblogs data, that includes large data sizes and high velocity, has motivated researchers to develop new techniques for data management and analysis on microblogs.

In this tutorial, we aim to provide a comprehensive review for almost all existing techniques and systems for microblogs data management and analysis, since the inception of Twitter in 2006. Figure 1 depicts a summary of the techniques and systems that will be covered in this tutorial in a timeline format. The horizontal axis in Figure 1 represents the year of publication or system release for each technique/system, while the vertical axis represents the research topic. The techniques are then classified into three categories: (1) techniques that deal with real-time data, i.e., very recent data, depicted by a filled black circle, (2) techniques that deal with historical data, depicted by a blank circle, and (3) techniques that deal with both real-time and historical data, depicted by a blank triangle.

Figure 2 gives the detailed tutorial outline and timing that lasts for **1.5 hours**. The first 10 minutes will be basically an introduction to the world of microblogs, a motivation for the need for research in data management and analysis, and a description of the tutorial outline with the aid of Figure 1. Following that quick introduction and motivation, the rest of the tutorial is divided into the following three parts: (1) microblogs data analysis, (2) microblogs data management, and (3) microblogs systems. The following sections describe the contents and scope of each part.

This work is partially supported by the National Science Foundation, USA, under Grants IIS-1525953, CNS-1512877, IIS-0952977 and IIS-1218168 and the University of Minnesota Doctoral Dissertation Fellowship.

II. Part I: Microblogs Data Analysis

This part of the tutorial covers existing work for microblogs data analysis that are depicted in the first six rows of Figure 1. As shown in Figure 2, this part will take 30 minutes in total, as five minutes for each of the following six types of analysis:

- **Event detection and analysis** [1], [11], [36], [38], [42], [44]: This work exploits the fact that microblogs users post many updates on on-going events. Such updates are identified, grouped, and analyzed to either discover events in real time [1], [36] or analyze long term events [11], [38], [42], [44], e.g., elections.
- **Recommendation** [7], [14], [32]: This work exploits microblogs user-generated contents as means for catching user preferences. The main objective is to recommend real-time news to read [32], authentic and credible users to follow [7], or users who share similar interests [14].
- **Sentiment and semantic analysis** [3], [28], [30]: This work quantifies positive and negative opinions in social media discussions as a pre-processing step for this data to be used in other analysis tasks, e.g., product review or election candidate surveying.
- **Visual analysis** [12], [15], [21], [27], [37], [42]: This work either explores visual analysis by itself [12], [15] or associate it with other analysis tasks [21], [27], [37], [42] to visually explore the output. In general, different types of graphs and maps are used visually to summarize the large amounts of microblogs data and its contents.
- **User analysis** [14], [18], [22], [45]: This work is mainly interested in analyzing user information to either identify

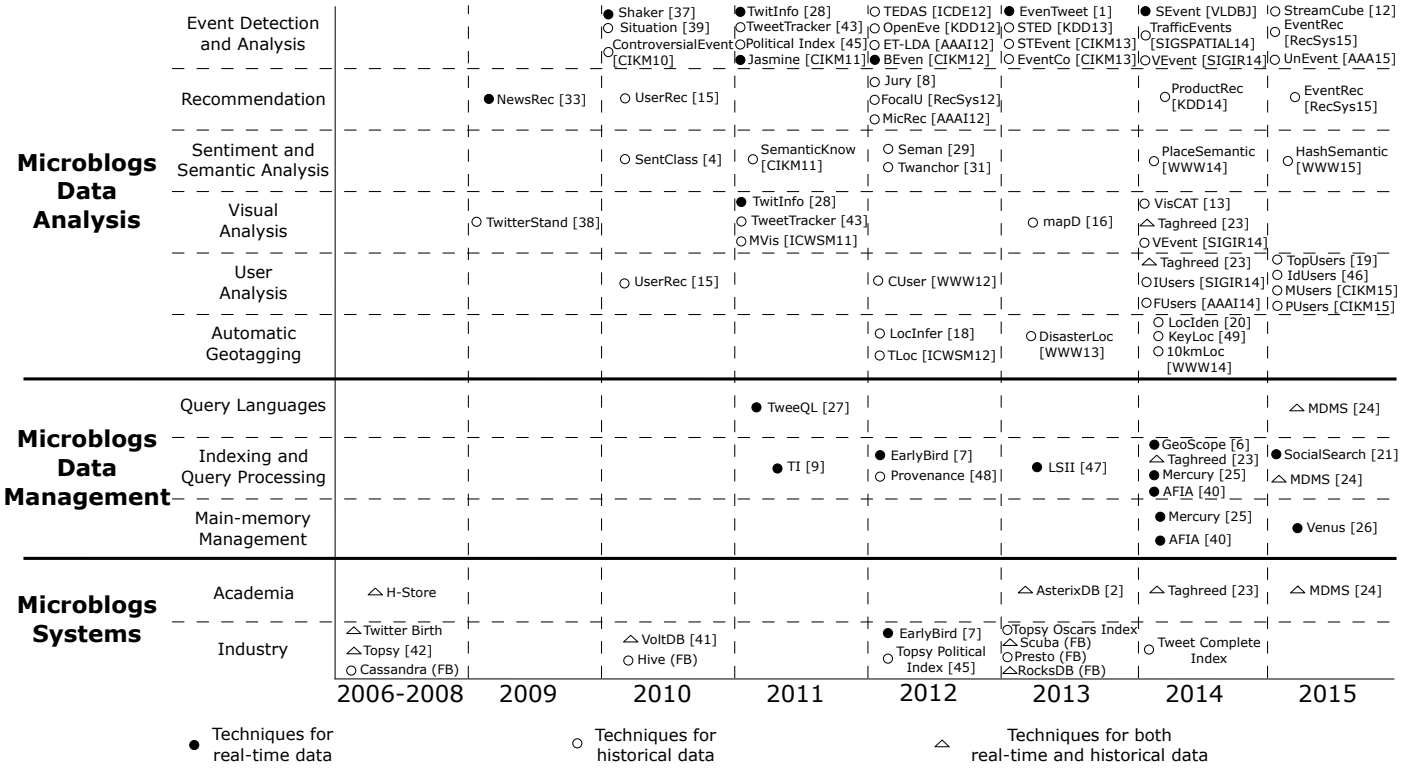

Fig. 1. Microblogs Timeline

top influential users in certain regions or topics [18], [22], [45], or discover users with similar interests [14]. Such users, or group of users, can be used in several applications, including posting ads and enhancing their social graph.

- **Automatic geotagging** [17], [19], [35], [48]: This work tries to attach more location information with microblogs data based on analyzing their contents. This is mainly motivated by the small percentage of geotagged microblogs (only 2% of tweets) that is faced by the need of many location-aware applications that can be built on top of microblogs [24], [25].

Other analysis tasks are addressed on microblogs data include news extraction [32], [37], topic extraction [16], [34], geo-targeted advertising [31], and generic social media analysis [41], [49]. However, for a limited time budget and audience engagement, we outline the main analysis tasks that span a wide variety of interests and applications.

III. Part II: Microblogs Data Management

This part of the tutorial covers existing work for microblogs data management that are depicted in the seventh to ninth rows of Figure 1. As shown in Figure 2, this part will take 30 minutes in total, and include the following three data management areas:

- **Query languages** [23], [26]: This work provides generic query languages that support SQL-like queries on top of microblogs. This facilitates basic Select-Project-Join operators that are able to support a variety of queries, either on top of Twitter APIs [26] or in core systems that supports microblogs data management [23].

- **Indexing and query processing**: This work includes various indexing techniques that have been proposed to index incoming microblogs either in memory [5], [6], [22], [24], [39], [46], [47] or in disk [8], [22]. This includes keyword search based on temporal ranking [6], [8], single-attribute search based on generic ranking functions [46], spatial-aware search that exploits location information in microblogs [24], and personalized social-aware search that exploits the social graph and produces user-specific search results [20]. Each search technique comes with its index and query processor, and tackles certain limitations in its preceding techniques.

- **Main-memory management** [24], [25], [39]: This work includes techniques that optimize for main-memory consumption and utilization. In recent microblogs indexing and query processing techniques, almost all of them depend on main-memory to host microblogs in their index structures. Thus, some techniques are equipped for main-memory management such that memory resources are efficiently utilized, either for aggregate queries [39] or basic search queries that retrieve individual data items [24], [25].

Introduction and Background (10 minutes)

- Importance and applications of microblogs
- Research literature outline

Part I: Microblogs Data Analysis (30 minutes)

- Event detection and analysis (5 minutes)
- Recommendation (5 minutes)
- Sentiment and semantic analysis (5 minutes)
- Visual analysis (5 minutes)
- User analysis (5 minutes)
- Automatic geotagging (5 minutes)

Part II: Microblogs Data Management (30 minutes)

- Query languages (5 minutes)
- Indexing and query processing (20 minutes)
  - Keyword search (5 minutes)
  - Generic search (5 minutes)
  - Spatial-aware search (5 minutes)
  - Social-aware search (5 minutes)
- Main-memory management (5 minutes)

Part III: Microblogs Systems (20 minutes)

- Challenges and motivational case studies (5 minutes)
- Microblogs in existing big data systems (15 minutes)
  - Twitter: Digesting, indexing, and querying tweets
  - AsterixDB: Persisting fast data
  - VoltDB: Transactions on fast data
  - Taghreed: Querying, analyzing, and visualizing microblogs

Fig. 2. Detailed Outline and Timing of **1.5 hours** Tutorial

IV. Part III: Microblogs Systems

This part highlights the current state and the challenges of managing microblogs data through existing big data systems [2], [6], [29], [22], [40], depicted in the last two rows in Figure 1. As shown in Figure 2, this part will take 20 minutes in total and will give a briefing on system challenges and motivational case studies. Then, we highlights the data management technologies in the following four systems:

- **Twitter** [6], [29]: Twitter is the most famous microblogging service provider worldwide. Every now and then, Twitter reveals technical details about its underlying systems components. We give a comprehensive review about the data management aspects of Twitter systems, per their published papers [6], [29].
- **AsterixDB** [2]: AsterixDB is a generic big data management system that can support various data sources. Recently, AsterixDB has extended its components to support fast data [13], e.g., microblogs, natively in the system. We review the fast data support in AsterixDB, which shows the challenges in persisting fast data in existing systems.
- **VoltDB** [40]: VoltDB is an emerging big data management system that is mainly optimized for database transactions on fast data, e.g., microblogs. We review the challenges of supporting transactional applications on fast data and solutions at the system level.
- **Taghreed** [22]: Taghreed is the first end-to-end holistic system to support microblogs data management. It supports different types of queries on both very recent data as well as historical data. We review the main components of Taghreed system highlighting its distinguishing properties to handle microblogs data.

V. Intended Audience

This tutorial targets researchers and developers who are interested in analyzing, processing, and building applications on microblogs data. Our tutorial highlights the major analysis tasks that have been conducted on microblogs, along with a road map for their different challenges and approaches. In addition, we introduce core data management techniques and holistic systems that support efficient and scalable querying on microblogs, highlighting challenges and opportunities in this area. Hence, attending this tutorial would help the audience to get more familiar with the state-of-the-art research on microblogs and identify the possible opportunities for future work.

VI. Presenters Biography

Amr Magdy is a fifth year Ph.D. student at the Department of Computer Science and Engineering, University of Minnesota. He received his M.Sc. at the same department in 2013. His main research interest is microblogs data management. He has been selected as a finalist for Microsoft Research PhD Fellowship 2014-2016 for his work in microblogs research and has been awarded the University of Minnesota Doctoral Dissertation Fellowship in 2015 for his dissertation focus on microblogs. His research work on microblogs has been incubated by Bing GeoSpatial team and has been selected among best papers in ICDE 2014. During his PhD, he has collaborated with Microsoft Research in Redmond in joint research on microblogs from which they got three publications and working on the fourth. Amr is also the main architect and developer for Taghreed system; the first holistic system for microblogs data management. Amr has published extensively in the area of microblogs data in top research venues, including IEEE ICDE, ACM SIGSPATIAL, IEEE TKDE, and IEEE MDM.

Mohamed F. Mokbel is an associate professor at University of Minnesota. His current research interests focus on database systems, GIS, and big spatial data. His research work has been recognized by five best paper awards and by the NSF CAREER award 2010. Mohamed is/was General co-chair of SSTD 2011, Program co-chair of ACM SIGSPATIAL GIS 2008-2010, and IEEE MDM 2011, 2014, Tutorial co-chair for ICDE 2014, and Sponsorship Co-Chair for ACM SIGMOD 2015-2016. He is in the editorial board of ACM Transactions on Database Systems, VLDB Journal, ACM Transactions on Spatial Algorithms and Systems, and GeoInformatica. Mohamed has held various visiting positions at Microsoft Research-Redmond and Hong Kong Polytechnic University. Mohamed is a founding member of ACM SIGSPATIAL and is currently serving as an elected chair of ACM SIGSPATIAL for the term 2014-2017. For more information, please visit: www.cs.umn.edu/∼mokbel

REFERENCES

[1] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. EvenTweet: Online Localized Event Detection from Twitter. In *VLDB*, 2013.

[2] Sattam Alsubaiee and et. al. AsterixDB: A Scalable, Open Source BDMS. *PVLDB*, 7(14), 2014.

[3] Adam Bermingham and Alan F. Smeaton. Classifying Sentiment in Microblogs: Is Brevity an Advantage? In *CIKM*, 2010.

[4] After Boston Explosions, People Rush to Twitter for Breaking News. http://www.latimes.com/business/technology/la-fi-tn-after-boston-explosions-people-rush-to-twitter-for-breaking-news-20130415,0,3729783.story, 2013.

[5] Ceren Budak, Theodore Georgiou, Divyakant Agrawal, and Amr El Abbadi. GeoScope: Online Detection of Geo-Correlated Information Trends in Social Networks. In *VLDB*, 2014.

[6] Michael Busch, Krishna Gade, Brian Larson, Patrick Lok, Samuel Luckenbill, and Jimmy Lin. Earlybird: Real-Time Search at Twitter. In *ICDE*, 2012.

[7] Caleb Chen Cao, Jieying She, Yongxin Tong, and Lei Chen. Whom to Ask? Jury Selection for Decision Making Tasks on Micro-blog Services. *PVLDB*, 5(11), 2012.

[8] Chun Chen, Feng Li, Beng Chin Ooi, and Sai Wu. TI: An Efficient Indexing Mechanism for Real-Time Search on Tweets. In *SIGMOD*, 2011.

[9] Sina Weibo, China Twitter, comes to rescue amid flooding in Beijing. http://thenextweb.com/asia/2012/07/23/sina-weibo-chinas-twitter-comes-to-rescue-amid-flooding-in-beijing/, 2012.

[10] Facebook Statistics. http://newsroom.fb.com/company-info/, 2015.

[11] Wei Feng, Jiawei Han, Jianyong Wang, Charu Aggarwal, and Jianbin Huang. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration Over the Twitter Stream. In *ICDE*, 2015.

[12] Thanaa Ghanem, Amr Magdy, Mashaal Musleh, Sohaib Ghani, and Mohamed Mokbel. VisCAT: Spatio-Temporal Visualization and Aggregation of Categorical Attributes in Twitter Data. In *SIGSPATIAL*, 2014.

[13] Raman Grover and Michael Carey. Data Ingestion in AsterixDB. In *EDBT*, 2015.

[14] John Hannon, Mike Bennett, and Barry Smyth. Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In *RecSys*, 2010.

[15] Harvard Tweet Map. worldmap.harvard.edu/tweetmap/, 2013.

[16] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering Geographical Topics In The Twitter Stream. In *WWW*, 2012.

[17] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location Inference Using Microblog Messages. In *WWW*, 2012.

[18] Jinling Jiang, Hua Lu, Bin Yang, and Bin Cui. Finding Top-k Local Users in Geo-Tagged Social Media Data. In *ICDE*, 2015.

[19] Guoliang Li, Jun Hu, Jianhua Feng, and Kian-Lee Tan. Effective Location Identification from Microblogs. In *ICDE*, 2014.

[20] Yuchen Li, Zhifeng Bao, Guoliang Li, and Kian-Lee Tan. Real Time Personalized Search on Social Networks. In *ICDE*, 2015.

[21] Amr Magdy, Louai Alarabi, Saif Al-Harthi, Mashaal Musleh, Thanaa Ghanem, Sohaib Ghani, Saleh Basalamah, and Mohamed Mokbel. Demonstration of Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *ICDE*, 2015.

[22] Amr Magdy, Louai Alarabi, Saif Al-Harthi, Mashaal Musleh, Thanaa Ghanem, Sohaib Ghani, and Mohamed Mokbel. Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *SIGSPATIAL*, 2014.

[23] Amr Magdy and Mohamed Mokbel. Towards a Microblogs Data Management System. In *MDM*, 2015.

[24] Amr Magdy, Mohamed F. Mokbel, Sameh Elnikety, Suman Nath, and Yuxiong He. Mercury: A Memory-Constrained Spatio-temporal Real-time Search on Microblogs. In *ICDE*, 2014.

[25] Amr Magdy, Mohamed F. Mokbel, Sameh Elnikety, Suman Nath, and Yuxiong He. Venus: Scalable Real-time Spatial Queries on Microblogs with Adaptive Load Shedding. *TKDE*, 2015.

[26] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Tweets as Data: Demonstration of TweeQL and TwitInfo. In *SIGMOD*, 2011.

[27] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*, 2011.

[28] Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. Adding Semantics to Microblog Posts. In *WSDM*, 2012.

[29] Gilad Mishne, Jeff Dalton, Zhenghua Li, Aneesh Sharma, and Jimmy Lin. Fast Data in the Era of Big Data: Twitter's Real-time Related Query Suggestion Architecture. In *SIGMOD*, 2013.

[30] Gilad Mishne and Jimmy Lin. Twanchor Text: A Preliminary Study of the Value of Tweets as Anchor Text. In *SIGIR*, 2012.

[31] New Enhanced Geo-targeting for Marketers. https://blog.twitter.com/2012/new-enhanced-geo-targeting-for-marketers.

[32] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using Twitter to Recommend Real-Time Topical News. In *RecSys*, 2009.

[33] Public Health Emergency, Department of Health and Human Services. http://nowtrending.hhs.gov/, 2015.

[34] Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. Characterizing Microblogs with Topic Models. In *ICWSM*, 2010.

[35] Christian Safran, Victor Manuel García-Barrios, and Martin Ebner. The Benefits of Geo-Tagging and Microblogging in m-Learning: a Use Case. In *MindTrek*, 2009.

[36] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *WWW*, 2010.

[37] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: News in Tweets. In *GIS*, 2009.

[38] Vivek K. Singh, Mingyan Gao, and Ramesh Jain. Situation Detection and Control using Spatio-temporal Analysis of Microblogs. In *WWW*, 2010.

[39] Anders Skovsgaard, Darius Sidlauskas, and Christian S. Jensen. Scalable Top-k Spatio-temporal Term Querying. In *ICDE*, 2014.

[40] Michael Stonebraker and Ariel Weisberg. The VoltDB Main Memory DBMS. *IEEE Data Engineering Bulletin*, 36(2):21–27, 2013.

[41] Topsy Analytics: Find the insights that matter. www.topsy.com, 2014.

[42] TweetTracker: track, analyze, and understand activity on Twitter. tweettracker.fulton.asu.edu/, 2014.

[43] Twitter Statistics. https://about.twitter.com/company, 2015.

[44] Topsy Analytics for Twitter Political Index. topsylabs.com/election/.

[45] Norases Vesdapunt and Hector Garcia-Molina. Identifying Users in Social Networks with Limited Information. In *ICDE*, 2015.

[46] Lingkun Wu, Wenqing Lin, Xiaokui Xiao, and Yabo Xu. LSII: An Indexing Structure for Exact Real-Time Search on Microblogs. In *ICDE*, 2013.

[47] Junjie Yao, Bin Cui, Zijun Xue, and Qingyun Liu. Provenance-based Indexing Support in Micro-blog Platforms. In *ICDE*, 2012.

[48] Rong Zhang, Xiaofeng He, Aoying Zhou, and Chaofeng Sha. Identification of Key Locations based on Online Social Network Activity. In *ASONAM*, 2014.

[49] Junzhou Zhao, John C.S. Lui, Don Towsley, Pinghui Wang, and Xiaohong Guan. Sampling Design on Hybrid Social-Affiliation Networks. In *ICDE*, 2015.