# Microblogs Data Management Systems: Querying, Analysis, and Visualization

Mohamed F. Mokbel Amr Magdy

Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN, USA
{mokbel,amr}@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 analyzing events and users activities and critical applications like discovering health issues and rescue services. Consequently, major research efforts are spent to manage, analyze, and visualize microblogs data to support different applications. In this tutorial, we give a 1.5 hours overview about microblogs data management, analysis, visualization, and systems. The tutorial gives a comprehensive review for research on core data management components to support microblogs queries at scale. This includes system-level issues and on-going work on supporting microblogs data through the rising wave of big data systems. In addition, the tutorial reviews research on microblogs data analysis and visualization. Through its different parts, the tutorial highlights the challenges and opportunities in microblogs data research.

## 1. 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 [30]. The distinguished nature of microblogs data, that includes large data sizes and high velocity, has motivated researchers to develop new techniques for data management to support microblogs analysis and visualization at scale.

This research is capitally supported by NSF grants IIS-0952977, IIS-1218168, IIS-1525953, CNS-1512877, and the University of Minnesota Doctoral Dissertation Fellowship.



*SIGMOD '16, June 26–July 1, 2016, San Francisco, CA, USA.*



DOI: http://dx.doi.org/10.1145/2882903.2912570

In this tutorial, we aim to provide a comprehensive review for almost all existing techniques and systems for microblogs data management, analysis, and visualization, 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 five minutes will be an introduction to the world of microblogs and a motivation for the need for research in data management. Following that quick introduction, the rest of the tutorial is divided into four parts: (1) microblogs data analysis (20 minutes), (2) microblogs data management (30 minutes), (3) microblogs data visualization (15 minutes), and (4) microblogs systems (20 minutes). The following sections describe the contents and scope of each part.

## 2. PART 1: MICROBLOGS DATA ANALYSIS

This part covers the literature for microblogs data analysis that are depicted in the first four rows of Figure 1. As shown in Figure 2, this part will take 20 minutes in total, where five minutes will be dedicated for each of the following four 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 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 tasks is to recommend real-time news to read [32], authority users to follow [7], or users who share similar interests [14].
- **Sentiment and semantic analysis** [3, 27, 29]: 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.
- **User analysis** [14, 17, 20, 45]: This work is mainly interested in analyzing user information to identify top influential

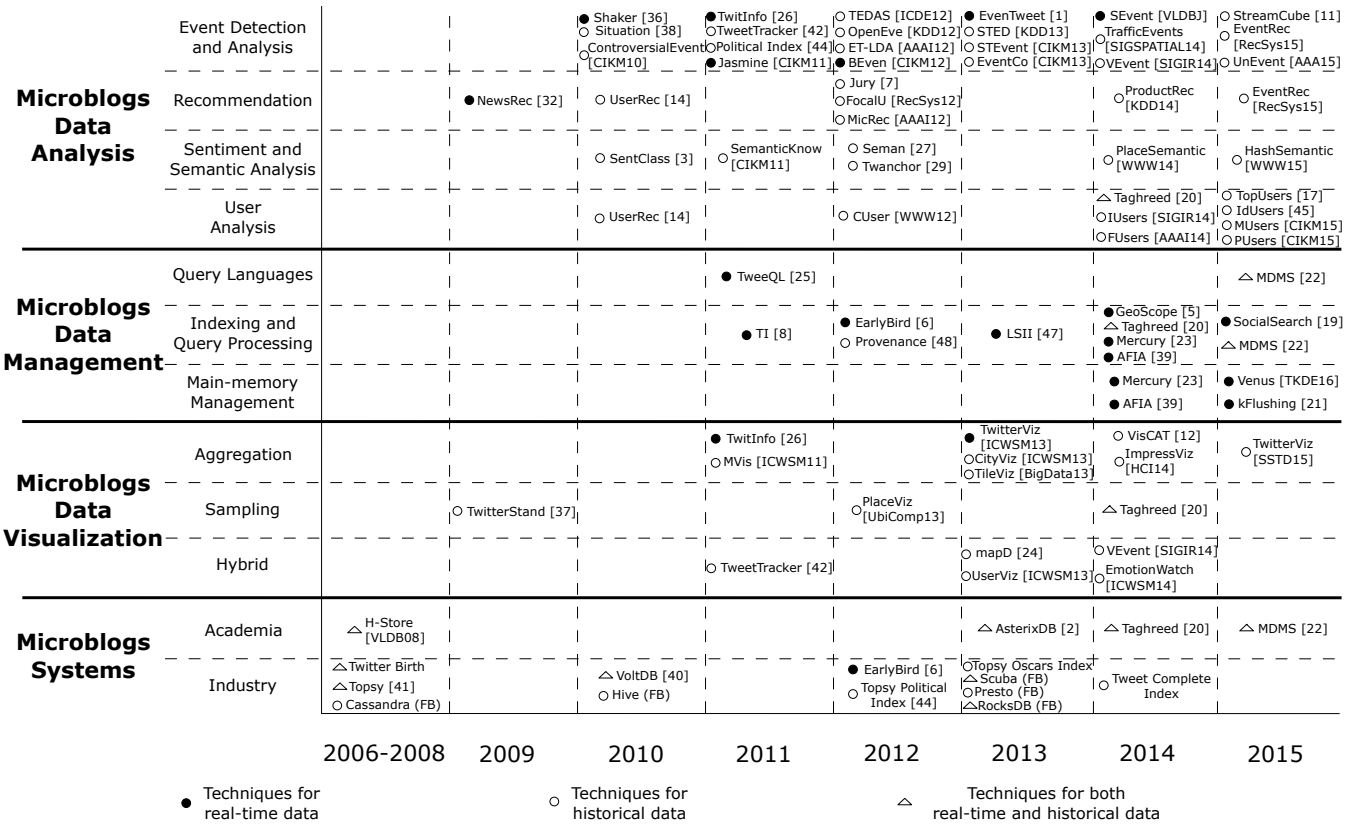


**Figure 1: Microblogs Timeline**

users in certain regions or topics [17, 20, 45], or discover users with similar interests [14]. Such users, or group of users, can be used in several scenarios, including posting ads and enhancing their social graph.

Other analysis tasks are addressed on microblogs data in both academic community, e.g., news extraction [32, 37], topic extraction [15, 34], and automatic geotagging [16, 18, 35, 49], and industrial community, e.g., geo-targeted advertising [30] and generic social media analysis [41, 50]. However, for limited time budget and audience engagement, we outline the main performed analysis that span a wide variety of interests and applications.

# 3. PART 2: MICROBLOGS DATA MANAGEMENT

This part of the tutorial covers existing work for microblogs data management that are depicted in the fifth to seventh 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** [22, 25]: 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 [25] or in core systems that supports microblogs data management [22].

- **Indexing and query processing**: This work includes various indexing techniques that have been proposed to index incoming microblogs either in memory [5, 6, 20, 23, 39, 47, 48] or in disk [8, 20]. This includes keyword search based on temporal ranking [6, 8], single-attribute search based on generic ranking functions [47], spatial-aware search that exploits location information in microblogs [23], personalized social-aware search that exploits the social graph and produces user-specific search results [19], and aggregate queries [5, 39] that find frequent keywords and correlated locations instead of individual microblog items. Each search technique comes with its index and query processor, and tackles certain limitations in its preceding techniques.

- **Main-memory management** [21, 23, 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 [21, 23].

# 4. PART 3: MICROBLOGS DATA VISUALIZATION

This part covers existing work for microblogs data visualization that are depicted in eighth to tenth rows of Figure 1. As shown in Figure 2, this part will take 15 minutes in total, and divides visualization techniques in microblogs into three areas:

- **Aggregation-based Visualization** [12, 26, 46]: This work overcomes the difficulty of visualizing tremendous amount

**Introduction and Background (5 minutes)**

- Importance and applications of microblogs
- Research literature outline

**Part 1: Microblogs Data Analysis (20 minutes)**

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

**Part 2: 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 3: Microblogs Data Visualization (15 minutes)**

- Aggregation-based Visualization (5 minutes)
- Sampling-based Visualization (5 minutes)
- Hybrid Visualization (5 minutes)

**Part 4: 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

**Figure 2: Detailed Outline and Timing of 1.5 hours Tutorial**

of data through showing only aggregate information that summarize the large amounts of microblogs data and its contents. Such aggregation is application-dependent and is usually performed based on major attributes, like temporal aggregation [46], spatial aggregation [12, 46], or keyword aggregation [26].

- **Sampling-based Visualization** [20, 37]: This work selects only a sample of microblogs and visualize them to users. The sample of data is of interest to the application, which can be selected based on some relevance criteria, e.g., tweets with news stories [37], or based on interest in certain attribute(s) [20].

- **Hybrid Visualization** [24, 42]: This work combines both aggregation and sampling in visualizing microblogs of interest. The two steps are either used simultaneously with the purpose of interest in certain data items, e.g., analyzing certain events [42], or used sequentially [24, 31] with purpose of bounding the number of microblogs to visualize.

# 5. PART 4: MICROBLOGS SYSTEMS

This part highlights the current state and the challenges of managing microblogs data through existing big data systems [2, 6, 28, 20, 40], depicted in the last two rows in Figure 1. As shown in Figure 2, this part will take 20 minutes 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, 28]: 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, 28].

- **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** [20]: 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.

# 6. INTENDED AUDIENCE

This tutorial targets researchers and developers who are interested in analyzing, processing, and building applications on microblogs data. Our tutorial gives extensive survey on core components that manages microblogs data at scale, highlighting challenges and opportunities in this area. In addition, we highlight the major analysis tasks that have been conducted on microblogs, along with detailed explanation for different visualization techniques that are used to show their output. 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.

# 7. PREVIOUS TUTORIALS

Mohamed Mokbel and Amr Magdy have a similar 90-minute tutorial to be presented in IEEE International Conference on Data Engineering (ICDE) on May 2016. There are two main differences between the two tutorials: (1) Part 3 of this tutorial (visualization) is completely new and was not proposed nor will be presented in the ICDE tutorial. (2) While the ICDE tutorial mainly focus on the data analysis part, this tutorial gives less focus on data analysis in favor of the data management part and the new visualization part.

# 8. 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] H. Abdelhaq, C. Sengstock, and M. Gertz. EvenTweet: Online Localized Event Detection from Twitter. In *VLDB*, 2013.

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

[3] A. Bermingham and A. 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] C. Budak, T. Georgiou, D. Agrawal, and A. E. Abbadi. GeoScope: Online Detection of Geo-Correlated Information Trends in Social Networks. In *VLDB*, 2014.

[6] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin. Earlybird: Real-Time Search at Twitter. In *ICDE*, 2012.

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

[8] C. Chen, F. Li, B. C. Ooi, and S. 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] W. Feng, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration Over the Twitter Stream. In *ICDE*, 2015.

[12] T. Ghanem, A. Magdy, M. Musleh, S. Ghani, and M. Mokbel. VisCAT: Spatio-Temporal Visualization and Aggregation of Categorical Attributes in Twitter Data. In *SIGSPATIAL*, 2014.

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

[14] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In *RecSys*, 2010.

[15] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering Geographical Topics In The Twitter Stream. In *WWW*, 2012.

[16] Y. Ikawa, M. Enoki, and M. Tatsubori. Location Inference Using Microblog Messages. In *WWW*, 2012.

[17] J. Jiang, H. Lu, B. Yang, and B. Cui. Finding Top-k Local Users in Geo-Tagged Social Media Data. In *ICDE*, 2015.

[18] G. Li, J. Hu, J. Feng, and K. Tan. Effective Location Identification from Microblogs. In *ICDE*, 2014.

[19] Y. Li, Z. Bao, G. Li, and K.-L. Tan. Real Time Personalized Search on Social Networks. In *ICDE*, 2015.

[20] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T. Ghanem, S. Ghani, and M. Mokbel. Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *SIGSPATIAL*, 2014.

[21] A. Magdy, R. Alghamdi, and M. F. Mokbel. On Main-memory Flushing in Microblogs Data Management Systems. In *ICDE*, 2016.

[22] A. Magdy and M. Mokbel. Towards a Microblogs Data Management System. In *MDM*, 2015.

[23] A. Magdy, M. F. Mokbel, S. Elnikety, S. Nath, and Y. He. Mercury: A Memory-Constrained Spatio-temporal Real-time Search on Microblogs. In *ICDE*, 2014.

[24] MapD. http://www.mapd.com/, 2013.

[25] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Tweets as Data: Demonstration of TweeQL and TwitInfo. In *SIGMOD*, 2011.

[26] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *CHI*, 2011.

[27] E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In *WSDM*, 2012.

[28] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin. Fast Data in the Era of Big Data: Twitter's Real-time Related Query Suggestion Architecture. In *SIGMOD*, 2013.

[29] G. Mishne and J. Lin. Twanchor Text: A Preliminary Study of the Value of Tweets as Anchor Text. In *SIGIR*, 2012.

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

[31] One Million Tweet Map. http://onemilliontweetmap.com/, 2016.

[32] O. Phelan, K. McCarthy, and B. 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] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing Microblogs with Topic Models. In *ICWSM*, 2010.

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

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

[37] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *GIS*, 2009.

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

[39] A. Skovsgaard, D. Sidlauskas, and C. S. Jensen. Scalable Top-k Spatio-temporal Term Querying. In *ICDE*, 2014.

[40] M. Stonebraker and A. 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] N. Vesdapunt and H. Garcia-Molina. Identifying Users in Social Networks with Limited Information. In *ICDE*, 2015.

[46] I. Weber and V. R. K. Garimella. Visualizing User-Defined, Discriminative Geo-Temporal Twitter Activity. In *ICWSM*, 2014.

[47] L. Wu, W. Lin, X. Xiao, and Y. Xu. LSII: An Indexing Structure for Exact Real-Time Search on Microblogs. In *ICDE*, 2013.

[48] J. Yao, B. Cui, Z. Xue, and Q. Liu. Provenance-based Indexing Support in Micro-blog Platforms. In *ICDE*, 2012.

[49] R. Zhang, X. He, A. Zhou, and C. Sha. Identification of Key Locations based on Online Social Network Activity. In *ASONAM*, 2014.

[50] J. Zhao, J. C. Lui, D. Towsley, P. Wang, and X. Guan. Sampling Design on Hybrid Social-Affiliation Networks. In *ICDE*, 2015.