Microblogs: A Renewable Spatio-temporal Fortune

Amr Magdy[1,2]
[1]Department of Computer Science and Engineering
[2]Center for Geospatial Sciences
University of California, Riverside

Abstract

Web users are long-standing sources for rich renewable datasets that are exploited in a wide variety of applications. Such datasets include significant spatial and temporal challenges that shape today's techniques and future technologies in the spatial community. This article highlights microblogs as a renewable source of user-generated data with a great fortune of spatial and temporal information about users, locations, and events that are exploited in rich applications. The articles covers both data management and analysis, discussing some of the existing challenges and future directions.

1 Introduction

Microblogs data is a form of web-based user-generated data that is micro in length and could be generated in various formats, e.g., micro text such as tweets, reviews, or comments, or micro activities such as location check-ins. This data is very easy and quick for users to generate, consequently, it became very popular as a communication method among hundreds of millions of web users on various microblogging platforms [36, 120]. The majority of these users are using microblogs from mobile devices around the clock, which has brought location, time, and mobility as core aspects of microblogs data research and applications. Furthermore, the richness of microblogs content has empowered their popularity in various applications. Microblogs contain news, updates on on-going events, reviews, location information, language information, user information, discussions in politics, products, and many others. This is being exploited in a wide variety of practical applications [82, 91], including public health [98, 115], disaster response [22, 23, 35, 53, 54, 59, 65], public safety [113], education [132], real-time news delivery [16], geo-targeted advertising [93], and several disciplines of academic research such as social science [119], information modeling [56], human dynamics [111], engagement in education [117], political sciences [118], behavioral sciences [110], and even medical-related research [116].

In this article, we highlight the spatial aspects of microblogs data research and analysis applications. The research literature of microblogs is rich and includes several major research communities, e.g., data management, natural language processing, and information retrieval. However, this article mostly focuses on the data management community, from a spatial perspective, that provides scalable infrastructures for indexing and querying microblogs. In addition, we highlight major research topics that exploit data management infrastructures to build spatial applications and analysis modules on top of microblogs, such as visual analysis, user analysis queries, and event detection. Other major research directions, e.g., natural language processing and information retrieval, have dedicated survey papers to review parts of their literature [26, 41].

The rest of this article is organized as follows. Section 2 highlights major data management techniques that deal with spatio-temporal microblogs. Section 3 highlights major data analysis research that builds on top of data management infrastructures to analyze microblogs. Finally, Section 4 concludes the article and discusses challenges and future directions.

2 Microblogs Data Management

The data management stack on microblogs started from the high-level declarative query languages, where there are few attempts in the literature to standardize query languages tailored for the needs of microblogs, and inspired by SQL query language: TWEEQL [88] and MQL [81, 83]. In addition to the standard *Select-Project-Join* constructs, both languages support spatial and temporal components that shows the inherent spatio-temporal nature of microblogs data. TWEEQL [88] supports built-in spatial and temporal filters to enable users to manipulate spatial and temporal attributes natively. Moreover, it allows user-defined functions that are also used to perform higher-level spatio-temporal analysis, such as automatic geotagging and chronological tracking. MQL [81, 83], on another hand, promotes the *temporal* aspect as a mandatory in all queries, while spatial attributes are manipulated through native constructs in both continuous and snapshot queries.

A core component of managing microblogs data at large scale is spatio-temporal indexing and query processing techniques. The research literature has techniques for both aggregate and non-aggregate queries. All existing queries on microblog include both temporal and top-k aspects regardless their other details. This is attributed to the nature of microblogs as they come in large numbers around the clock. This large number mandates retrieving the most useful k microblogs based on a certain top-k ranking function, otherwise, many useless data will be reported. In addition, being a kind of streaming data, the data is real-time by nature and many users and applications are interested in recent microblogs. This inspired almost all the existing techniques to embed the time aspect by default in the query signature, unless it is disabled by the user. In fact, without using the time aspect, a query might retrieve data from several years ago, which leads to a significant querying overhead. So, by disabling this default option, users become aware of the implications on querying performance if they consider data of long temporal periods.

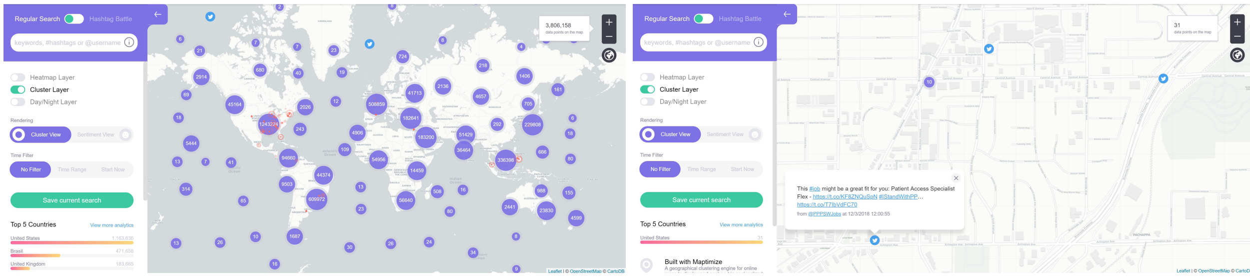
A generic query signature that represents the major existing microblogs queries is: "*Find top-k microblogs/keywords ranked based on a ranking function F.*" In non-aggregate queries that retrieve individual microblogs, the ranking function *F* can be *temporal* [4, 18, 21], *spatio-temporal* [84, 85], *significance-temporal* [129], or *socio-temporal* [37]. In aggregate queries [17, 62, 80, 109], the spatial and temporal aspects are used as filters for queried data and the ranking functions mostly depend on keyword counts and their derived measure, e.g., trendline slope, except GEOSCOPE [17] that employs a correlation measure.

Almost all indexing techniques of microblogs are optimized for high digestion rates in a main-memory index for real-time data indexing, and secondary storage indexing is assumed to have older data to query historical microblogs. The only exception is TI [21] that primarily uses a disk-based index. With such inherent real-time nature, all major index structures are temporal in a sense [4, 18, 19, 38, 72, 84, 129, 134] to support various temporal ranking functions that are listed above (JUDICIOUS [139] is an exception though). For spatial attributes, both aggregate and non-aggregate queries are supported on spatial microblogs. Several spatial index structures are designed for different aggregate queries [17, 62, 80, 109], and a few are designed for non-aggregate queries to retrieve individual microblogs within certain spatial range or nearby a certain location [81, 83, 84, 85]. Details about differences among these different indexing and query processing techniques can reviewed in [78].

3 Microblogs Data Analysis

This section highlights the major data analysis research for analysis tasks that provide high-level functionality on microblogs. As microblogs data analysis is a broad literature, we focus on novel research that is related to the data management community. Specifically, we focus on geovisualization, user analysis, event detection and analysis, and automatic geotagging.

Geovisualization. Spatial and temporal attributes are used heavily in visualizing microblogs data for different applications, e.g., political and disastrous event analysis, disease outbreaks detection, and user communities analysis. The challenges faced in visualizing microblogs data aligns with the general challenges in visualizing



(a) Aggregated microblogs on continent level (b) Individual microblogs on street level

Figure 1: Example of aggregation-based visualization based on the spatial dimension.

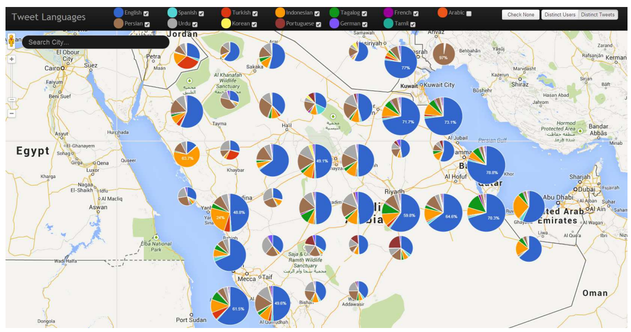


Figure 2: Example of aggregation-based visualization based on both spatial and language dimensions.

other types of big data [14, 20, 31, 32, 45, 67, 95]. So, several pieces of the proposed research for big data visualization can be used for microblogs data as one type of big datasets. However, we highlight how visualization work on microblogs, as user-generated data, promote temporal and spatial aspects in almost all use cases. With the micro-length content, microblogs are easy to be generated by users all the time, e.g., a user can easily generate a tweet in a few seconds or less. This leads to generating a large number of data records in relatively short times. Visualizing such large numbers is beyond the capacity of existing frontend technologies, such as mapping technologies, e.g., Google Maps. So, visualization techniques that focus on microblogs try to address this problem by either aggregation, sampling, or a combination of both. So, we classify microblogs visualization literature into three categories of techniques: (1) aggregation-based techniques, (2) sampling-based techniques, or (3) hybrid techniques. The visualization modules in all these categories use underlying querying modules, both aggregate and non-aggregate queries, to retrieve the data to be visualized. Examples of each category are highlighted next.

Aggregation-based visualization techniques [3, 28, 39, 52, 58, 87, 101, 108, 112, 125, 128, 131, 136] reduce the amount of data to be visualized through visualizing aggregate summaries of microblogs at different levels of aggregation, e.g., different spatial levels or temporal levels, rather than visualizing individual microblogs. Such aggregation is application-dependent and is usually performed either based on major attributes, e.g., temporal aggregation [28, 52], spatial aggregation [39, 128], or keyword aggregation [28, 112], or based on derived attributes, e.g., sentiment [52, 101]. Thus, these techniques are lossless and present all available information in a summarized form without ignoring any portion of the data. Aggregation could be based on a single attribute (one-dimensional) or multiple attributes (multi-dimensional). Figure 1 shows an example of aggregation-based visualization based on a single attribute, the spatial attribute [94]. In Figure 1(a), spatial regions that have a large number of data points visualize a variable-size circle that shows the number of points in this region. On the contrary, regions that have sparse data, Arctic Ocean and Norwegian Sea in Figure 1(a), visualize the actual data points. On zooming on the map view, more detailed data is visualized up to the street level that shows detailed

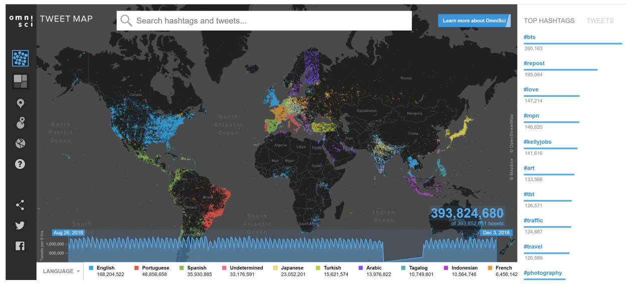


Figure 3: Example of sampling-based visualization for tweets with different languages in different locations.

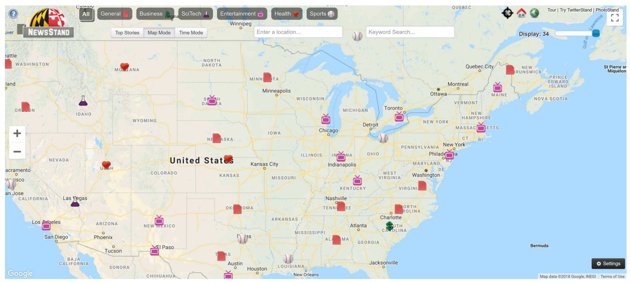


Figure 4: Example of sampling-based visualization for news tweets.

data points, as depicted in Figure 1(b) that shows street-level data in Riverside, California. Figure 2 shows an example of aggregation-based visualization based on two attributes, the spatial attribute and the language attribute [39]. In this case, number of microblogs is aggregated in each spatial region and the visualized circle categories data based on the language attribute to show percentage of microblogs posted in English, Arabic, Indonesian, Persian, etc.

Sampling-based visualization techniques [79, 107, 126] reduce the amount of visualized data through sampling. A sample of data is selected and visualized as a representative for the whole dataset, while the rest of data is not visualized. The sampling technique can be classified based on different dimensions. A sample could be a query-guided sample or an arbitrary sample. An example for a query-guided sample is OmniSci TweetMap[1] (Figure 3) that samples tweets based their language as the query predicate filters data based on the language attribute. Another example is TwitterStand [107] (Figure 4) that samples tweets based on textual content that have news stories. For certain queries, the query predicate is generating a lot of data that still cannot be visualized efficiently. In this case, applications, e.g., [79], selects an arbitrary data sample to reduce the data size. Unlike aggregation-based techniques that are lossless, sampling-based techniques might be lossy or lossless depending on the application and the size of query result. If certain application queries are generating a reasonable sample size, then all data points are considered. Otherwise, such as in arbitrary sampling, a subset of data points are ignored and the sampling is lossy.

Hybrid visualization techniques are used in several applications allow to use both aggregation and sampling to reduce the amount of data to be visualized [24, 60, 87, 89, 90, 114, 135]. For example, event analysis applications [89, 114] sample microblogs based on their relevance to specific events. Then, event data need to be aggregated to summarize the event highlights to users, e.g., showing changes over time, space, users, or topics. Such applications usually do not encounter challenges in visualizing their data as the data size is reduced over two different phases, sampling and aggregation, which leads to significant reduction in their size and ease the visualization task.

Automatic geotagging. Geo-locations are heavily exploited in several microblogs applications, such as

[1]https://www.omnisci.com/demos/tweetmap/

localized event detection [2], geo-targeted advertising [93], local news extraction [107], user interest inference [44], and finding local active users [61]. With all such importance of geo-location data in microblogs applications, still the majority of microblogs are not associated with precise location information. In fact, a small percentage ($< 4\%$) of popular microbologging data, e.g., Twitter, is associated with locations sourced from user devices. This triggered a need to associate location information with more microblogs data automatically to exploit as much microblogs as possible in location-aware applications. However, traditional geotagging techniques are limited for enriching microblogs location data due to the brevity of microblogs textual content. Such brief text contains a lot of abbreviations and noisy words that make it hard for named entity recognizers to extract accurate places and locations.

Most of microblogs geotagging techniques use a single microblog record at a time [70, 99, 104, 48, 55, 86, 73, 6, 30, 29]. This works on two stages. The first stage is a feature extraction stage, such as keywords and named entities places extraction from brief textual contents. The second stages is a classification based on the extracted features. A common problem in these techniques is that precision significantly drops for the practical margins of error distance, which represents the distance between actual location and predicted location. To overcome this problem, a newer technique [70] proposed to process microblogs as collections instead of individual records, which has shown tremendous enhancement in prediction precision and recall (95+%) within 100 meters error distance.

Event detection and analysis. Event detection and analysis has gained tremendous attention with the rise of microblogging platforms [1, 2, 5, 8, 9, 12, 25, 27, 34, 37, 43, 49, 50, 51, 63, 64, 68, 71, 74, 76, 92, 96, 97, 102, 103, 105, 127, 133, 137, 138, 142, 144, 146, 147, 148]. The reason is the popularity of event-related updates that are posted by users through microblogs around the clock. This includes a wide variety of both short-term and long-term events, such as concerts, crimes, sports matches, accidents, natural disasters, social unrest, festivals, traffic jams, elections, and conflicts. Analyzing the event-related microblogs enabled several applications at different levels of importance, including crucial applications, e.g., rescue services and emergency response [22, 23, 35, 53, 54, 59, 65], leisure applications, e.g., detecting entertainment events, and in-between applications, e.g., news extraction based on events [16], event-driven advertising [93], public opinion analysis for political campaigns [122, 123], and analyzing protests and social unrest [10, 57, 121].

The rich literature of event detection and analysis on microblogs can be categorized into three main categories: (1) detecting arbitrary events, (2) detecting specific types of events, e.g., earthquakes or sports games, and (3) analyzing pre-defined events. Detecting arbitrary events have either no predefined or at most very high-level characteristics. For example, looking for local events in a certain city [2, 37] without determining any specific characteristics of such events. Such techniques generally has some, or all, of five main stages: (a) *filtering & feature extraction*, (b) *grouping*, (c) *scoring*, (d) *summarization*, and (e) *visualization*. All these stages use spatial and temporal aspects of microblogs heavily [2, 8, 37, 64, 68, 102, 127, 137, 138, 142, 144, 148], due to the inherent nature of events that has spatial and temporal extents in most use cases. Only the summarization stage that uses the textual content more than other features.

Detecting event of specific types is another major direction of event detection techniques that have a set of distinguishing information to characterize the event, e.g., crimes, earthquakes, or traffic jams. These techniques have three main stages: (a) *feature extraction*, (b) *event classification*, and (c) *visualization*. The feature extraction mostly use temporal and textual attributes [51, 71, 74, 145], and occasionally uses spatial attributes [51]. The visualization stage though uses spatial and temporal aspects in almost all cases to provide users with a useful visualization of where and when real-life events happen. This also happens in event analysis applications, that do not detect events, instead they take an event as an input and analyze its data in different ways, including spatio-temporal visualization and slicing. An example is the Syrian revolution that is a long-term event that is known beforehand with a set of features such as keywords and locations. Such applications work mostly on two stages: (a) *filtering* and (b) *visualization & analysis*, where both of them heavily use spatial and temporal attributes [11, 114].

User analysis. The importance of microblogs in different applications originates from its user-generated nature, where hundreds of millions of users worldwide are posting around the clock. Among the major analysis directions is analyzing the user behavior related to different topics, locations, and communities based on their profiles and content of their microblogs. The literature of user analysis techniques on microblogs can be classified into techniques that either (1) find top-k users according to a certain ranking criteria [75, 61, 15, 124, 46, 106, 33, 77, 130, 7, 13, 140, 143, 69, 40], e.g., top-k influential users for a certain topic or top-k active users in a certain location, to provide useful answers for higher-level applications, or (2) classify users based on certain characteristics [47, 42, 66, 100, 141]. The first category, which is the majority of existing work, makes use of spatial and temporal attributes at different stages of the analysis. First, a user is profiled based on their spatial and temporal behavior, in addition to other pieces of information, such as keywords and followers/friends. Second, the user spatio-temporal information is used for indexing or modeling. Third, the indexed/modeled spatio-temporal signatures are used by query processors to answer application-level queries.

4 Conclusions and Future Directions

This article has highlighted the major spatio-temporal data management and analysis research work on user-generated microblogs as a renewable source of data that flows all the time from all locations. The rich literature of research on microblogs data faces several big challenges and is still rich with opportunities on different fronts. Focusing on spatio-temporal aspects, combing social information with real-time spatio-temporal information still underutilized in supporting scalable personalized queries on real-time data. In addition, there is almost no work on studying the implications of the novel real-time indexing techniques on query optimization models for spatio-temporal data.

In terms of data analysis, there are several untackled challenges. First, real-time geotagging of microblogs data. Although recent work started to tackle this problem [29, 30], there are still challenges in reducing the geotagging time due to the high computational cost of this task. Achieving the goal of attaching locations to microblgs as they come will widely impact a plethora of location-aware applications that are built on top of microblogs. Second, developing a unified event detection framework that allows users to express different types of event-based queries. Third, integrating the rich literature of user analysis techniques with the scalable data management infrastructures, e.g., indexes and query processors, in microblogs systems. Such integration will allow a variety of user-centric applications to be supported at scale.

Another direction that is not yet popular in geovisualization of micorblogs is adaptive sampling. Currently, any interaction for end users with the map view will not change the content of this sample. User interactions only change the subset of this sample that is shown on the map. On the contrary, interactive sampling changes the sample content based on user interactions. At the beginning, an initial sample of 100K, for example, is visualized from all locations including 30K microblogs in US. When the user filters out data to show only US microblogs, the visualized US microblogs can be increased to 100K as it is solely visualized. Such interactive technique is exploiting the whole capacity of frontend technologies while increasing the overall amount of data visualized to users. Such technique is not heavily used and has several research challenges to support large-scale data.

In addition to enhancing different analysis modules, there is a dire need to integrate such rich literature of analysis techniques with microblogs data systems to widen the impact of microblogs research in a practical sense. Such integration will have tremendous impact of a plethora of applications that benefit the society, the research community, and business applications, including public health, disaster response, public safety, and education. The feasible way to achieve such goal is abstracting different analysis tasks on microblogs into basic building blocks that can be supported in microblogs systems, inspired by SELECT-PROJECT-JOIN building blocks in SQL database management systems. Such task is huge and shall be started with developing generic frameworks for different analysis tasks, as discussed in our prior work [78].

Acknowledgments

This work is partially supported by the National Science Foundation, USA, under grants IIS-1849971, SES-1831615, and CNS-1837577.

References

[1] H. Abdelhaq, M. Gertz, and A. Armiti. Efficient Online Extraction of Keywords for Localized Events in Twitter. *GeoInformatica*, 21(2):365–388, 2017.

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

[3] Y. Abdelsadek, K. Chelghoum, F. Herrmann, I. Kacem, and B. Otjacques. Community Extraction and Visualization in Social Networks Applied to Twitter. *Information Sciences*, 424:204–223, 2018.

[4] M. K. Agarwal, D. Bansal, M. Garg, and K. Ramamritham. Keyword Search on Microblog Data Streams: Finding Contextual Messages in Real Time. In *EDBT*, 2016.

[5] M. Akbari, X. Hu, L. Nie, and T.-S. Chua. From Tweets to Wellness: Wellness Event Detection from Twitter Streams. In *AAAIz*, 2016.

[6] H. Al-Olimat, K. Thirunarayan, V. L. Shalin, and A. P. Sheth. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. In *COLING*, 2018.

[7] Z. Z. Alp and S. G. Ögüdücü. Influential User Detection on Twitter: Analyzing Effect of Focus Rate. In *ASONAM*, 2016.

[8] N. Alsaedi, P. Burnap, and O. Rana. Can We Predict A Riot? Disruptive Event Detection using Twitter. *ACM TOIT*, 17(2):18, 2017.

[9] N. Alsaedi, P. Burnap, and O. F. Rana. Automatic Summarization of Real World Events Using Twitter. In *ICWSM*, 2016.

[10] New Study Quantifies Use of Social Media in Arab Spring. www.washington.edu/news/2011/09/12/new-study-quantifies-use-of-social-media-in-arab-spring/, 2011.

[11] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal and Events based Analysis of Topic Popularity in Twitter. In *CIKM*, 2013.

[12] N. Avudaiappan, A. Herzog, S. Kadam, Y. Du, J. Thatche, and I. Safro. Detecting and Summarizing Emergent Events in Microblogs and Social Media Streams by Dynamic Centralities. In *IEEE Big Data*, 2017.

[13] M. Bartoletti, S. Lande, and A. Massa. Faderank: An Incremental Algorithm for Ranking Twitter Users. In *WISE*, 2016.

[14] L. Battle, R. Chang, and M. Stonebraker. Dynamic Prefetching of Data Tiles for Interactive Visualization. In *SIGMOD*, 2016.

[15] I. Bizid, N. Nayef, P. Boursier, S. Faïz, and A. Doucet. Identification of Microblogs Prominent Users during Events by Learning Temporal Sequences of Features. In *CIKM*, 2015.

[16] 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.

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

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

[19] H. Cai, Z. Huang, D. Srivastava, and Q. Zhang. Indexing Evolving Events from Tweet Streams. *TKDE*, 27(11):3001–3015, 2015.

[20] H. Chavan and M. F. Mokbel. Scout: A GPU-Aware System for Interactive Spatio-temporal Data Visualization. In *SIGMOD*, 2017.

[21] C. Chen, F. Li, B. C. Ooi, and S. Wu. TI: An Efficient Indexing Mechanism for Real-Time Search on Tweets. In *SIGMOD*, 2011.

[22] How Twitter, Facebook, WhatsApp And Other Social Networks Are Saving Lives During Disasters. http://www.huffingtonpost.in/2017/01/31/how-twitter-facebook-whatsapp-and-other-social-networks-are-sa_a_21703026/, 2017.

[23] 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.

[24] S. Counts and K. Fisher. Taking It All In? Visual Attention in Microblog Consumption. In *ICWSM*, 2011.

[25] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang. Discover Breaking Events with Popular Hashtags in Twitter. In *CIKM*, 2012.

[26] N. F. F. da Silva, L. F. S. Coletta, and E. R. Hruschka. A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised Learning. *ACM Computing Surveys*, 49(1):15:1–15:26, 2016.

[27] N. D. Doulamis, A. D. Doulamis, P. C. Kokkinos, and E. M. Varvarigos. Event Detection in Twitter Microblogging. *IEEE Transactions Cybernetics*, 46(12):2810–2824, 2016.

[28] C. Drescher, G. Wallner, S. Kriglstein, R. Sifa, A. Drachen, and M. Pohl. What Moves Players?: Visual Data Exploration of Twitter and Gameplay Data. In *CHI*, 2018.

[29] N. Duong-Trung, N. Schilling, and L. Schmidt-Thieme. Near Real-time Geolocation Prediction in Twitter Streams via Matrix Factorization based Regression. In *CIKM*, 2016.

[30] R. Dutt, K. Hiware, A. Ghosh, and R. Bhaskaran. SAVITR: A System for Real-time Location Extraction from Microblogs during Emergencies. *CoRR*, abs/1801.07757, 2018.

[31] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis. MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration. In *ICDE*, 2016.

[32] A. Eldawy, M. F. Mokbel, and C. Jonathan. HadoopViz: A MapReduce Framework for Extensible Visualization of Big Spatial Data. In *ICDE*, 2016.

[33] M. Enoki, Y. Ikawa, and R. Raymond. User Community Reconstruction Using Sampled Microblogging Data. In *WWW Companion*, 2012.

[34] A. E. Erdoğan, T. Yilmaz, O. C. Sert, M. Akyüz, T. Özyer, and R. Alhajj. From Social Media Analysis to Ubiquitous Event Monitoring: The Case of Turkish Tweets. In *ASONAM*, 2017.

[35] How Facebook Is Transforming Disaster Response. `https://www.wired.com/2016/11/facebook-disaster-response/`, 2016.

[36] Facebook Statistics. `http://newsroom.fb.com/company-info/`, 2018.

[37] W. Feng, C. Zhang, W. Zhang, 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.

[38] L. Gao, Y. Wang, D. Li, J. Shao, and J. Song. Real-time social media retrieval with spatial, temporal and social constraints. *Neurocomputing*, 253:77–88, 2017.

[39] 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.

[40] S. Ghosh, N. K. Sharma, F. Benevenuto, N. Ganguly, and P. K. Gummadi. Cognos: Crowdsourcing Search for Topic Experts in Microblogs. In *SIGIR*, 2012.

[41] A. Giachanou and F. Crestani. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys*, 49(2):28:1–28:41, 2016.

[42] Z. Gilani, E. Kochmar, and J. Crowcroft. Classification of Twitter Accounts into Automated Agents and Human Users. In *ASONAM*, 2017.

[43] H. Gu, M. Gartrell, L. Zhang, Q. Lv, and D. Grunwald. AnchorMF: Towards Effective Event Context Identification. In *CIKM*, 2013.

[44] L. Guo, D. Zhang, Y. Wang, H. Wu, B. Cui, and K.-L. Tan. Co 2: Inferring personal interests from raw footprints by connecting the offline world with the online world. *ACM Transactions on Information Systems (TOIS)*, 36(3):31, 2018.

[45] T. Guo, K. Feng, G. Cong, and Z. Bao. Efficient Selection of Geospatial Data on Maps for Interactive and Visualized Exploration. In *SIGMOD*, 2018.

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

[47] L. He and J. Luo. "What Makes a Pro Eating Disorder Hashtag": Using Hashtags to Identify Pro Eating Disorder Tumblr Posts and Twitter Users. In *IEEE Big Data*, 2016.

[48] B. J. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *CHI*, 2011.

[49] Y. Hu, A. John, F. Wang, and S. Kambhampati. ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback. In *AAAI*, volume 12, 2012.

[50] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan. STED: Semi-supervised Targeted-interest Event Detectionin in Twitter. In *SIGKDD*, 2013.

[51] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan. Automatic Targeted-domain Spatio-temporal Event Detection in Twitter. *GeoInformatica*, 20(4):765–795, 2016.

[52] R. B. Hubert, E. Estevez, A. G. Maguitman, and T. Janowski. Examining Government-Citizen Interactions on Twitter Using Visual and Sentiment Analysis. In *DG.O*, 2018.

[53] Hurricane Harvey Victims Turn to Twitter and Facebook. http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/, 2017.

[54] In Irma, Emergency Responders' New Tools: Twitter and Facebook. https://www.wsj.com/articles/for-hurricane-irma-information-officials-post-on-social-media-1505149661, 2017.

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

[56] Prediction, Optimization and Control for Information Propagation on Networks: A Differential Equation and Mass Transportation Based Approach. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1620342`, 2017.

[57] A Nobel Peace Prize for Twitter? `www.csmonitor.com/Commentary/Opinion/2009/0706/p09s02-coop.html`, 2009.

[58] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa. Visual Exploration of Changes in Passenger Flows and Tweets on Mega-City Metro Network. *IEEE Transactions on Big Data*, 2(1):85–99, 2016.

[59] Japan earthquake: how Twitter and Facebook helped.
http://www.telegraph.co.uk/technology/twitter/8379101/Japan-earthquake-how-Twitter-and-Facebook-helped.html, 2011.

[60] J. Jia, C. Li, X. Zhang, C. Li, M. J. Carey, and S. Su. Towards Interactive Analytics and Visualization on One Billion Tweets. In *SIGSPATIAL*, 2016.

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

[62] C. Jonathan, A. Magdy, M. F. Mokbel, and A. Jonathan. GARNET: A Holistic System Approach For Trending Queries in Microblogs. In *ICDE*, 2016.

[63] J. Kalyanam, S. Velupillai, M. Conway, and G. Lanckriet. From Event Detection to Storytelling on Microblogs. In *ASONAM*, 2016.

[64] T. Kaneko and K. Yanai. Visual Event Mining from the Twitter Stream. In *WWW Companion*, 2016.

[65] Embrace of Social Media Aids Flood Victims in Kashmir.
https://www.nytimes.com/2014/09/13/world/asia/embrace-of-social-media-aids-flood-victims-in-kashmir.html, 2014.

[66] A. Khatua and A. Khatua. Cricket World Cup 2015: Predicting User's Orientation through Mix Tweets on Twitter Platform. In *ASONAM*, 2017.

[67] A. Kim, E. Blais, A. G. Parameswaran, P. Indyk, S. Madden, and R. Rubinfeld. Rapid Sampling for Visualizations with Ordering Guarantees. *PVLDB*, 8(5):521–532, 2015.

[68] J. Krumm and E. Horvitz. Eyewitness: Identifying Local Events via Space-time Signals in Twitter Feeds. In *SIGSPATIAL*, 2015.

[69] P. Lahoti, G. D. F. Morales, and A. Gionis. Finding Topical Experts in Twitter via Query-dependent Personalized PageRank. In *ASONAM*, 2017.

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

[71] R. Li, K. H. Lei, R. Khadiwala, and K. C. Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *ICDE*, 2012.

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

[73] J. Lingad, S. Karimi, and J. Yin. Location Extraction From Disaster-related Microblogs. In *WWW*, 2013.

[74] M. Liu, K. Fu, C. Lu, G. Chen, and H. Wang. A Search and Summary Application for Traffic Events Detection Based on Twitter Data. In *SIGSPATIAL*, 2014.

[75] N. Liu, L. Li, G. Xu, and Z. Yang. Identifying Domain-Dependent Influential Microblog Users: A Post-Feature Based Approach. In *AAAI*, 2014.

[76] X. Liu, Q. Li, A. Nourbakhsh, R. Fang, M. Thomas, K. Anderson, R. Kociuba, M. Vedder, S. Pomerville, R. Wudali, et al. Reuters Tracer: A Large Scale System of Detecting & Verifying Real-time News Events from Twitter. In *CIKM*, 2016.

[77] X. Lu, P. Li, H. Ma, S. Wang, A. Xu, and B. Wang. Computing and Applying Topic-level User Interactions in Microblog Recommendation. In *SIGIR*, 2014.

[78] A. Magdy, L. Abdelhafeez, Y. Kang, E. Ong, and M. F. Mokbel. Microblogs Data Management: A Survey. *VLDB Journal*, 29(1):177–216, 2020.

[79] 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.

[80] A. Magdy, A. M. Aly, M. F. Mokbel, S. Elnikety, Y. He, S. Nath, and W. G. Aref. GeoTrend: Spatial Trending Queries on Real-time Microblogs. In *SIGSPATIAL*, 2016.

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

[82] A. Magdy and M. Mokbel. Microblogs Data Management and Analysis (Tutorial). In *ICDE*, 2016.

[83] A. Magdy and M. Mokbel. Demonstration of Kite: A Scalable System for Microblogs Data Management. In *ICDE*, 2017.

[84] 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.

[85] A. Magdy, M. F. Mokbel, S. Elnikety, S. Nath, and Y. He. Venus: Scalable Real-time Spatial Queries on Microblogs with Adaptive Load Shedding. *TKDE*, 28(2), 2016.

[86] J. Mahmud, J. Nichols, and C. Drews. Where Is This Tweet From? Inferring Home Locations of Twitter Users. In *ICWSM*, 2012.

[87] R. Makki, E. J. de Carvalho, A. J. Soto, S. Brooks, M. C. F. de Oliveira, E. E. Milios, and R. Minghim. ATR-Vis: Visual and Interactive Information Retrieval for Parliamentary Discussions in Twitter. *TKDD*, 12(1):3:1–3:33, 2018.

[88] 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.

[89] 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.

[90] A. J. McMinn, D. Tsvetkov, T. Yordanov, A. Patterson, R. Szk, J. A. Rodriguez Perez, and J. M. Jose. An Interactive Interface for Visualizing Events on Twitter. In *SIGIR*, 2014.

[91] M. Mokbel and A. Magdy. Microblogs Data Management Systems: Querying, Analysis, and Visualization (Tutorial). In *SIGMOD*, 2016.

[92] L. Mu, P. Jin, L. Zheng, E.-H. Chen, and L. Yue. Lifecycle-Based Event Detection from Microblogs. In *WWW Companion*, 2018.

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

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

[95] Y. Park, M. J. Cafarella, and B. Mozafari. Visualization-aware Sampling for Very Large Databases. In *ICDE*, 2016.

[96] M. Peng, J. Zhu, H. Wang, X. Li, Y. Zhang, X. Zhang, and G. Tian. Mining Event-Oriented Topics in Microblog Stream with Unsupervised Multi-View Hierarchical Embedding. *TKDD*, 12(3):38, 2018.

[97] A. Popescu and M. Pennacchiotti. Detecting Controversial Events from Twitter. In *CIKM*, 2010.

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

[99] Y. Qian, J. Tang, Z. Yang, B. Huang, W. Wei, and K. M. Carley. A Probabilistic Framework for Location Inference from Social Media. *CoRR*, abs/1702.07281, 2017.

[100] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. M. Jr. Characterizing and Detecting Hateful Users on Twitter. *CoRR*, abs/1803.08977, 2018.

[101] R. A. Rios, P. A. Pagliosa, R. P. Ishii, and R. F. de Mello. TSViz: A Data Stream Architecture to Online Collect, Analyze, and Visualize Tweets. In *SAC*, 2017.

[102] A. Ritter, O. Etzioni, S. Clark, et al. Open Domain Event Extraction from Twitter. In *SIGKDD*, 2012.

[103] K. Rudra, P. Goyal, N. Ganguly, P. Mitra, and M. Imran. Identifying Sub-events and Summarizing Disaster-Related Information from Microblogs. In *SIGIR*, 2018.

[104] K. Ryoo and S. Moon. Inferring Twitter User Locations with 10 km Accuracy. In *WWW Companion*, 2014.

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

[106] J. Sang, D. Lu, and C. Xu. A Probabilistic Framework for Temporal User Modeling on Microblogs. In *CIKM*, 2015.

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

[108] B. Sijtsma, P. Qvarfordt, and F. Chen. Tweetviz: Visualizing Tweets for Business Intelligence. In *SIGIR*, 2016.

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

[110] Understanding Social and Geographical Disparities in Disaster Resilience Through the Use of Social Media. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1620451`, 2017.

[111] Sparsity Models for Forecasting Spatio-Temporal Human Dynamics. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1737770`, 2017.

[112] D. Sundararaman and S. Srinivasan. Twigraph: Discovering and Visualizing Influential Words Between Twitter Profiles. In *Social Informatics*, 2017.

[113] J. G. Turet and A. P. C. S. Costa. Big Data Analytics to Improve the Decision-Making Process in Public Safety: A Case Study in Northeast Brazil. In *Springer ICDSST*, 2018.

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

[115] Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014. `https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6332a1.htm`, 2014.

[116] Harvard Medical School Researchers Awarded Twitter Data Grant. https://hms.harvard.edu/news/harvard-medical-school-researchers-awarded-twitter-data-grant, 2014.

[117] Twitter and Informal Science Learning and Engagement. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1438898`, 2017.

[118] The Power of Images: A Computational Investigation of Political Mobilization via Social Media. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1727459`, 2017.

[119] Twitter Data Changing Future of Population Research. http://news.psu.edu/story/474782/2017/07/17/research/twitter-data-changing-future-population-research, 2017.

[120] Twitter Statistics. `https://about.twitter.com/company`, 2018.

[121] The Twitter War: Social Media's Role in Ukraine Unrest. news.nationalgeographic.com/news/2014/05/140510-ukraine-odessa-russia-kiev-twitter-world/, 2014.

[122] Twitter a Big Winner in 2012 Presidential Election. www.computerworld.com/article/2493332/social-media/twitter-a-big-winner-in-2012-presidential-election.html, 2012.

[123] Topsy Analytics for Twitter Political Index. `https://blog.twitter.com/official/en_us/a/2012/a-new-barometer-for-the-election.html`.

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

[125] S. Wakamiya, A. Jatowt, Y. Kawai, and T. Akiyama. Analyzing Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs. In *WWW Companion*, 2016.

[126] Y. Wang, P. Siriaraya, Y. Nakaoka, H. Sakata, Y. Kawai, and T. Akiyama. A Twitter-Based Culture Visualization System by Analyzing Multilingual Geo-Tagged Tweets. In *ICADL*, 2018.

[127] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *CIKM*, 2011.

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

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

[130] S. Wu, L. Gong, W. Rand, and L. Raschid. Making Recommendations in a Microblog to Improve the Impact of a Focal User. In *RecSys*, 2012.

[131] X. Wu, L. Bartram, and C. Shaw. Plexus: An Interactive Visualization Tool for Analyzing Public Emotions from Twitter Data. *CoRR*, abs/1701.06270, 2017.

[132] Y. Wu. Language E-learning based on Learning Analytics in Big Data Era. In *International Conference on Big Data and Education*, 2018.

[133] C. Xing, Y. Wang, J. Liu, Y. Huang, and W.-Y. Ma. Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter. In *AAAI*, pages 2666–2672, 2016.

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

[135] Z. Yu, Z. Wang, L. Chen, B. Guo, and W. Li. Featuring, Detecting, and Visualizing Human Sentiment in Chinese Micro-Blog. *TKDD*, 10(4):48, 2016.

[136] M. A. Zayer and M. H. Gunes. Analyzing the Use of Twitter to Disseminate Visual Impairments Awareness Information. In *ASONAM*, 2017.

[137] C. Zhang, L. Liu, D. Lei, Q. Yuan, H. Zhuang, T. Hanratty, and J. Han. Triovecevent: Embedding-based Online Local Event Detection in Geo-tagged Tweet Streams. In *SIGKDD*, 2017.

[138] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han. Geoburst: Real-time Local Event Detection in Geo-tagged Tweet Streams. In *SIGIR*, 2016.

[139] D. Zhang, L. Nie, H. Luan, K.-L. Tan, T.-S. Chua, and H. T. Shen. Compact Indexing and Judicious Searching for Billion-Scale Microblog Retrieval. *ACM TOIS*, 35(3):27, 2017.

[140] J. Zhang, R. Zhang, J. Sun, Y. Zhang, and C. Zhang. TrueTop: A Sybil-Resilient System for User Influence Measurement on Twitter. *IEEE/ACM TON*, 24(5):2834–2846, 2016.

[141] Y. Zhang, Y. Fan, Y. Ye, X. Li, and E. Winstanley. Utilizing Social Media to Combat Opioid Addiction Epidemic: Automatic Detection of Opioid Users from Twitter. In *AAAI Workshops*, 2018.

[142] Y. Zhang, C. Szabo, Q. Z. Sheng, and X. S. Fang. SNAF: Observation Filtering and Location Inference for Event Monitoring on Twitter. *WWW Journal*, 21(2):311–343, 2018.

[143] J. Zhao, X. Gui, and F. Tian. A New Method of Identifying Influential Users in the Micro-Blog Networks. *IEEE Access*, 5:3008–3015, 2017.

[144] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Online Spatial Event Forecasting in Microblogs. *ACM TSAS*, 2(4):15, 2016.

[145] X. Zheng, A. Sun, S. Wang, and J. Han. Semi-Supervised Event-related Tweet Identification with Dynamic Keyword Generation. In *CIKM*, 2017.

[146] D. Zhou, L. Chen, and Y. He. An Unsupervised Framework of Exploring Events on Twitter: Filtering, Extraction and Categorization. In *AAAI*, 2015.

[147] D. Zhou, T. Gao, and Y. He. Jointly Event Extraction and Visualization on Twitter via Probabilistic Modelling. In *ACL*, volume 1, 2016.

[148] X. Zhou and L. Chen. Event Detection over Twitter Social Media Streams. *PVLDB*, 23(3):381–400, 2014.