Microblogs Content Management, Retrieval, Analysis and Visualization

Louai Alarabi

Computer Science Department, Umm Al-Qura University, SA

Summary

The popularity of microblogs data along with the easiness of filing and posting microblogs leads to unprecedented rates of data flows. In this study we explore various research directions of microblogging. We have examined the infrastructure of major microblogging systems and highlight the challenges of data analysis associated with microblogs. In our paper, we identified several gaps and research needed since all the previous research studies are not automated especially with the fast evolution of users' usage behaviors and trends. An automated management system and would be needed to cope up with the fast changes.

Key words:
Microblogs, Distributed Systems, Indexing, Querying, Information retrieval.

1. Introduction

Microblogs e.g., tweets [33], Facebook comments [6], and Foursquare check-ins [8] are among the most popular web services nowadays. For example, Twitter has more than 140 Million active users who generate more than 340 Million tweets [34], while Facebook has more than 950 Million users who post more than 3.2 Billion daily comments [7]. This tremendous amount of data posted every single second carries a broad spectrum of information, news, question/answers, personal social interactions, and other content types that a web user posts on their profiles. Motivated by the unprecedented flow of information produced from microblogs, recent research efforts have been conducted to manage and analyze microblogs data and discover and visualize events and trends from the micro blogosphere.

Recent research on microblogs mostly focused on keyword search in microblogs [2, 4, 17, 23], trend/event detection from microblogs [22, 30, 32, 36], understanding the social phenomena [11, 13], and microblogs ranking [5, 35]. In addition, locations of microblog entries have been used for either visualizing microblogs (e.g., tweet messages) on a geographical map [19, 20, 21, 31] or discovering localized events [10, 36]. This article is trying to provide a road map with a high-level review for recent research in microblogs.

2. Road Map

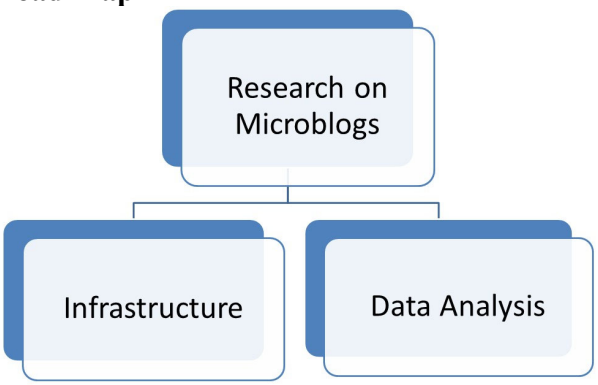

Fig. 1: Microblogs Research Map - Main Categories

Due to its popular and widespread use, several research efforts have started to explore various research directions related to microblogs. Such efforts can be categorized into two main categories as shown in Figure 1:

(1) Infrastructure, where the main focus is to build an infrastructure stack that is capable of providing the basic functionality of microblog systems. This goes along the way of the system stack starting from the underlying logging techniques [14] to large-scale machine learning techniques embedded in the Hadoop environment [16] to indexing microblog posts [2, 4, 37] for supporting keyword search queries through special APIs [23], and finally to designing a SQL-like query language interface for querying microblogs [19].

(2) Data analysis, where the main focus is to analyze microblog posts' contents, retrieved through specialized APIs, to get useful information. This includes semantic and sentiment analysis of microblog posts [1, 24, 26], decision-making [3], event and trend detection [15, 22, 30, 32], understanding the characteristics of microblog posts [17, 29], search queries [17], ranking microblog posts either based on their recency [5] or user profiles [35], and following users a recommendation [9] or news to read [28].

Manuscript received December 5, 2020
Manuscript revised December 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.12.3

Most recently, new research efforts have started to exploit the location information attached to microblog posts. Such efforts are either concerned with (a) Visualization, where each microblog post is plotted on a map based on its location [20, 21, 31], as a means of visual data analysis, (b) Automatic geotagging, where geotags are extracted from the contents of each microblog post to be used later in local event detections [36], or (c) offline analysis, that model the relationship between user interests, locations, and posted topics [10].

Microblog posts with location information can be considered as a spatio-temporal stream with very high arrival rates. There exist a wide-spectrum of related work over the last decade in the context of moving objects and traffic-related applications [12, 18, 25, 27, 38]. The main focus of such related work was on continuous queries that have to be registered first in the system, and then the query answer is collected later over time by the incoming spatiotemporal stream. This is mainly due to the underlying application (i.e., moving objects) that produces update streams of previous location entries. However, the model of the moving object does not fit well to handle microblogs. On the one hand, microblogs are flowing in the form of append-only streams with no updates. Besides, continuous queries are not of interest in the context of microblogs. Thus, in our review, we are not going to focus on either research on the moving objects model or its continuous queries along with the research map of microblogs.

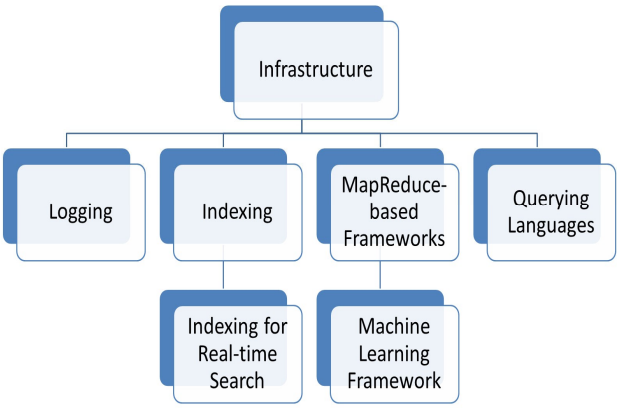

Fig.2: Microblogs Research Map - Infrastructure

3. MICROBLOGS INFRASTRUCTURE

As a newly rising web community that attracts several hundreds of millions of users and generates billions of data records daily, microblogs data, as any other kind of data, needs tools that provide the basic functionalities that are required to make use of the tremendous amount of microblogging data and hence provide applications on top of them. Providing efficient and scalable data management and retrieval methods is one of the most important tools that provide the most basic functionalities on microblogs data; storing, indexing, and querying. These tasks are among the most challenging tasks in microblogs.

Microblogs are flowing continuously around the world (and thus around the o'clock) at high arrival rates. This makes storing the entire microblogs stream is not a straight forward task even with the dropping storage cost. Hence, a system that deals with microblogs need to store a subset of this huge stream in a way that serves its purpose. Also, indexing such kind of data is challenging and has different aspects to consider. In addition to indexing on keywords, which is still very important, like any other type of data, microblogs services are coming with additional very important temporal and spatial dimensions. Chronological order is the default order for current microblogs interfaces, e.g., Twitter. These interfaces, along with the high arrival rates, enabled microblogs to provide real-time interactions of news and event discovery, as we are going to elaborate later in this article. Thus, the temporal dimension is considered the most important aspect of microblogs and hence needs to be primarily considered in any indexing technique for microblogs data. In addition, combined with the advances in wireless communication and GPS-equipped handheld devices, microblogs have entered a new era where locations can be attached to each posted microblog. Added location information gives extra information about either the whereabouts of the microblog issuer or the microblog contents. For example, Facebook added the option of "nearby" where users can state the nearby location of their status messages, Twitter automatically captures the GPS coordinates, if turned on, from mobile devices, while Foursquare is a microblogging service that is all around the location information and the whereabouts of its users. Thus, the location dimensions are again an important aspect of microblogs that should be considered in indexing microblogs data and provide a variety of applications and services.

In brief, indexing microblogs is a challenging task and has different aspects to consider; most importantly: keywords, temporal, and spatial dimensions. Again, querying the huge amount of microblogs data is a challenging research task that is studied carefully in the literature. To this end, in this section, we discuss these challenging research tasks on microblogs. Figure 2 shows the different subcategories of infrastructure research work.

Twitter APIs provide search services, including keywords and spatial services, on tweets. The research aspects of the Earlybird system that empowers indexing and

querying on Twitter data are published in [2]. Although the search services are distributed on many servers, the paper discusses the best we can get from a single box. Fundamentally, the Earlybird system focus on indexing tweets for real-time search. Earlybird servers provide results to blenders servers that apply additional filters and provide the user with the final results. The main contribution of this work is to enable real-time search on tweets within 10 seconds from tweet creation with an average query latency of 50 ms. One main challenge is to support inverted indexing for real-time search. Earlybird uses multiple index segments; each of them holds 8 million tweets. The newly arriving tweets fill segments sequentially, so only one segment is modified while the others are read-only. Each segment employs a hash-based dictionary that stored a list of postings with each keyword entry. The postings list is expanded dynamically based on four sizes that increase exponentially to efficiently handle both short and long lists of postings for different keywords. Another big challenge is to provide concurrency on the real-time index. The authors employ a single-writer multiple-readers algorithm that enables concurrent query evaluation while reducing the index update overhead.

The paper discusses some open problems and challenges:

1. Provide relevancy raking for real-time search results to eliminate noisy tweets, i.e., find the "popular" or "relevant" tweets instead of the most recent.
2. Provide personalized search results.
3. Making use of Twitter linkage with the other web documents, images, web URLs...etc., in real-time search results.

Before Earlybird details are revealed [2], TI [4] was the first trial to provide efficient real-time search on microblogs data. TI (which stands for Twitter Index) has chosen to handle the huge amount of social microblogs by indexing distinguished tweets immediately while postponing non-distinguished ones to later stages that are performed as periodic offline indexing. The important tweets are determined based on their contribution to answers a previously posted query. In other words, if any tweet can be included in the answer of any previously posted query to TI, then it is considered important and indexed immediately. Otherwise, it is buffered to a queue that is wiped periodically, and its contents are indexed in offline processing. As TI uses the in-memory index, like most of the systems that serve highly dynamic microblogs, when the memory constraints of TI are reached, it sheds less frequent topics (or groups of tweets) that did not receive enough queries for a long time (where the deletion time period is a system parameter that trades the system overhead with the query answers availability). This optimizes the system storage based on the incoming query load.

TweeQL [21] represents the earliest trial to standardize searching on microblogs. This work proposes a SQL-like language that works as a wrapper for Twitter APIs. This language enables its users to create streams of tweets that are actually sub-streams from the Twitter mainstream that satisfy certain predicates. For example, create a stream of tweets that contain the word "COVID-19" and issued no older than seven days ago. TweeQL provides flexible access to knowledge and information that are expressed implicitly and explicitly in Twitter data. Among the main research challenges addressed by TweeQL is uncertain selectivity. Some predicates, like location-based predicates, have no precise information to be evaluated against. In that sense, the WHERE statement of TweeQL needs to perform an uncertain condition evaluation. In addition, hitting Twitter APIs is expensive and encounter high latency. Thus, wrapping this, and providing more complex operation on top of it, needs an optimization plan to handle high latencies. TweeQL provides a query optimizer that generates a query plan to reduce the evaluation overhead. A third challenge is the huge amount of data that comes to the social stream. To handle such a high rate of tweets, TweeQL employs different components like a sampler, rate limiter, and latency enforcer.

McCullough et al. [23] proposed a methodology that evaluates the quality of search systems on microblogs. The main concept introduced in this work is essential tweets. Particularly, each search query results should contain its essential tweets that are defined as tweets with the exact match of query keywords sorted in the reverse temporal order. This was basically part of TREC's efforts to formalize studying Twitter data and provide standard data and methods to evaluate different systems and algorithms. This methodology succeeded in being applied for a different algorithm in the literature.

4. MICROBLOGS CONTENT ANALYSIS

The popularity of microblogs data along with the easiness of filing and posting microblogs leads to unprecedented rates of data flows, e.g., the current average Twitter rate is 4000 tweets/second. These microblogs' messages carry other news, information, knowledge, and social interactions. Thus, microblogs urge many research efforts to exploit such rich contents to accomplish different tasks, including semantic and sentiment analysis [1, 24, 26], decision making [3], event and trend detection [15, 22, 30, 32], ranking microblog posts based on their user profiles [35], recommending users to follow [9] or news to read [28], visual data analysis [20, 21, 31], automatic geotagging [36],

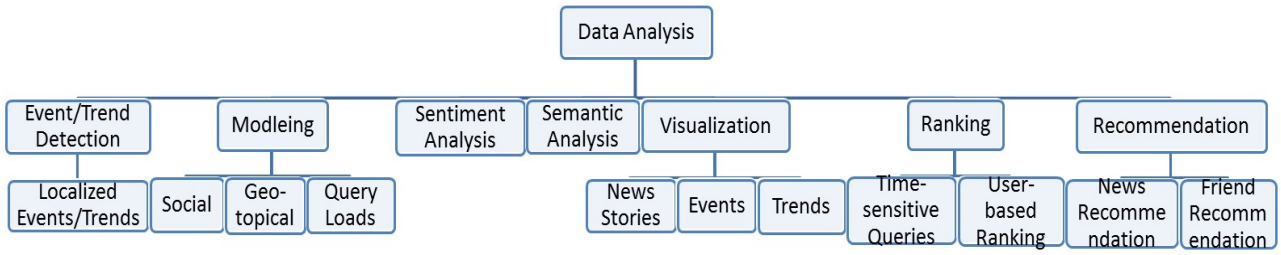

Fig. 3: Microblogs Research Map - Data Analysis

or user/topic/location modeling [10]. In this section, we summarize the proposed technique to use microblogs contents to accomplish these different tasks. Figure 3 shows the different subcategories of microblogs analysis research work.

News Extraction. The popularity of microblogging service nowadays made it a very rich, flexible, and fast media to propagate news stories. For example, the death of Michael Jackson is first tweeted one hour before any newsagent [31]. TwitterStand [31] is a system that tries to exploit news stories in Twitter data to provide users with a fast and rich news navigation experience. The system employs a set of tools and a dataset that enables the extraction of location information from tweets. It continuously crawls tweets from a pre-defined set of users who are known to post news stories. The system automatically identifies junk tweets, and if the tweet is news, it assigns a class (or classes) to each tweet, e.g., sport, science...etc. The user is then provided with a map-based user interface, provided with different filtering options with both topics and location filters that enable to navigate through the extracted news stories.

Phelan et al. [28] exploit news in tweets in a different way. They used online flowing tweets to recommend real-time topical news. In other words, when a new potential topical new story appears in real-time tweets, e.g., the death of Michael Jackson, it recommends to its users to read about this new topic. The main algorithm uses tweet content to score each tweet for being a news item or not.

Ranking. With the plethora of microblogs available to users, automatically selecting relevant microblogs comes as a high research priority so that microblogs data become more useful for users. [35] proposes a ranking technique that works in two directions: (i) identify the probability with which a certain user will retweet a certain tweet. In other words, given a list of tweets and a certain user, their approach is going to rank the tweet list based on the probability this user is expected to retweet them (ii) and the other direction rank a list of users based on their probability to retweet a certain tweet. Ranking incoming tweets depends on four features (i) Author-based: how the author of the tweet is elite, (ii) Tweet-based: the syntax of the tweet, (iii) Content-based, (iv) and User-based: how the user would rate the tweet. Secondly, ranking targeted users in which it gives whether this tweet is might interest the target user. In other words, a specific tweet should be sent to those who are strongly willing to retweet it. The algorithms' results are analyzed with a real user study.

Dong et al. [5] use Twitter data for ranking in a different sense. They use URLs and social graph of Twitter to improve search results for recency sensitive queries. The main idea is that real-time tweets contain fresh URLs that are not yet crawled by search engines. Thus, exploiting these URLs should improve results for queries that need hot and fresh results, e.g., real-time news search. In addition, they exploit the social relationships in the Twitter graph, i.e., following-follower relationships, to personalize the ranking score for each person instead of having just a general ranking for all. One main challenge that faces this system is the big portion of spam URLs posted on Twitter. So, the system first applies two filtering rules that can exclude a huge amount of spam URLs. The system then performs feature extraction from the remaining tweets to feed them into an existing raking algorithm that orders the tweets based on their importance.

Hannon et al. [9] perform another type of ranking, for users, not tweets, in the form of recommendation. The work aims at recommending new friends on Twitter social media based on content analysis and collaborative filtering techniques. In addition to demonstrating the potential of the real-time web and microblogging services to serve as a useful source of recommendation information, it examines how existing recommendation strategies can be usefully harnessed to solve important challenges following recommendation in the case of the real-time web.

Trend/event detection. One of the most currently appealing applications on top of microblogs is trend detection. In natural disasters, e.g., earthquakes, popular events, e.g., football games, celebrations...etc., it is popular nowadays that users post related microblogs, e.g., tweets with related hashtags. This enables the application to perform trend and event detection from microblogs data. With a huge amount of posted microblogs every hour, research efforts are made

to scale, discovering trends and events from microblogs. [32] proposes a system that detects and track situation and events from microblogs data. The system mainly classifies the incoming tweets into relevant or irrelevant to the topic, e.g., flu spread, earthquake damages...etc. The system uses spatial and temporal information included in the tweet besides the user settings. In [22], authors propose TwitterMonitior, a real-time monitoring tool that monitors Twitter stream and detects real-time events. TwitterMonitior mainly uses word counts to detect frequent keywords and group them to report events and trends.

TwitInfo [20] is one of the major research efforts in monitoring events over the social streams (that is also presented along with TweeQL in [21]). The system simplifies the problem statement to "identifying peak activities of the certain event" instead of discovering events from nothingness. System users register for monitoring events with specific keywords. From this moment on, the system shows peak activities for these specific keywords. Before describing how these peaks are detected, we want to highlight that this approach has two main advantages: (i) it personalizes the search results to have what interests them regardless of the other events that may be of less interest (ii) it does not need to store the whole data, instead only tweets for already registered queries need to be processed which reduce the system overhead significantly. To detect event peak activities, TwitInfo adapts algorithms from signal processing literature. The algorithm uses a time-based histogram and stores the frequency of each keyword in each histogram bin. When the system encounters a bin (or a group of consecutive bins) that has above-average frequencies, then this is identified as a peak for this event. By controlling the bin width, e.g., 1 minute, the system controls the detection smoothness and speed.

Jasmine's [36] system demonstration tends to narrow its scope to discover localized events that happen in a certain area of space. The main challenge for real-time local event detection is the lack of geotags in microblogs data. Jasmine proposed a new algorithm that extracts and attach location information with microblogs based on pre-loaded location entities data. This algorithm tags around 80% of the Twitter dataset used for their evaluation. Based on the attached location information, Jasmine cluster tweets based on closeness in time and space. Hence, each group of localized recent tweets represents an event that is presented to its users.

Sakaki et al. [30] added the notion of semantic analysis and event tracking to event detection from the Twitter stream. Dealing with Twitter users as social sensors, the system takes the meaning that users intend to express in their considering in analyzing the tweets for detecting related events. The system basically aims to discover real-time events, e.g., earthquakes, and predicate their trajectories from location information of tweets before they actually propagate on the physical surface of the earth. The event registers keywords of certain events, e.g., earthquake, and then classifies the incoming tweets into either "relevant" or "non-relevant" using an SVM classifier based on features that represent keywords, tweet-length...etc. The system successfully predicted the paths of Japan's latest earthquakes with 96% accuracy.

Modeling. Another type of analyzing microblogs data is modeling the relation between different aspects of the data. This kind of modeling is actually useful for several applications, including user profiling, content recommendation, and topic tracking. Also, the existence of the sheer amount of microblogs allow better modeling to answer questions like (i) How is information created and shared in different geographic locations? What is the inherent geographic variability of content? (ii) What are the spatial and linguistic characteristics of people? How does this vary across regions? (iii) What is a good model for human mobility? Can we discover patterns in users's usage of micro-blogging services? Authors in [10] developed a theoretical model that represents the relation between user interests, tweet content topics, and geographical information that are either attached or contained in the tweet. The model basically uses predefined attributes to make probabilistic decisions regarding either the topic or the location.

Lin and Mishne [17] tried to model the temporal evolution of data in the Twitter stream. In other words, they are trying to understand when the users begin to propagate a strong Twitter wave about a certain topic, what the authors call churn. The authors expected this to improve the quality of real-time search and help systems to provide users with enhanced results. This work models the changes in certain keywords over time using a frequency-based distribution (similar to the concept of histograms). Using this distribution, the deviations along the timeline represent how the event peak is changed, and hence how the users' interests are evolving over time. Analysis for a day over day showed that there is no cyclic pattern of churn. Another observation of the result is eliminating trend query reduce KL divergence. On the other hand, an hour over hour analysis showed a strong daily cyclic effect, removing trends actually increase churn and higher KL divergence values. Indeed, this shows that the typical life span of the trends is longer than one hour.

Ramage et al. [29] use tweet content to model the topics of the flowing real-time stream. The purpose of this is to enrich the user interface for microblogging service by attaching extra information with each microblog item that shows its

topical properties. In addition, this helps to develop tools to help Twitter users to discover new topics on the social stream, follow new users who post in related topics that interest them, and so on for the whole bunch of useful tools can be developed given the topical characteristics. The authors employ the LDA algorithm to discover the topical properties of tweets based on features extracted from its content.

Table 1: Limitations of Existing Systems

	Social Graph	Content	Timestamps	Location	Secondary Storage
Earlybird [2]	Not Used	Used	Used	Not Used	Not Used
TwitterMonitor [22]	Not Used	Used	Used	Not Used	Not Used
TweetRank [35]	Not Used	Used	Used	Not Used	Not Used
SituationDetector [32]	Not Used	Used	Used	Used	Not Used
TwitInfo [20]	Not Used	Used	Used	Not Used	Not Used
TweeQL [20]	Not Used	Used	Used	Used	Not Used
Jasmine [36]	Not Used	Used	Used	Used	Not Used
TwitterStand [31]	Not Used	Used	Used	Used	Not Used
GeoTopical Modeling [10]	Not Used	Used	Not Used	Used	Used
TI [4]	Not Used	Used	Used	Not Used	Used
Twitter Shaker [30]	Not Used	Used	Used	Used	Not Used
RecencyRanker [5]	Used	Used	Used	Not Used	Not Used
ChurnAnalyzer [17]	Not Used	Used	Used	Used	Used
TopicalAnalyzer [29]	Not Used	Used	Not Used	Not Used	Not Used
BloggerAnalyzer [11]	Used	Not Used	Not Used	Used	Used
NewsDecider [13]	Not Used	Used	Not Used	Not Used	Not Used
UserRecommender [9]	Used	Used	Not Used	Not Used	Not Used

Java et al. [11] try to model and understand the microblogging communities and how web users use this community in their lives. To this end, they study the topological and geographical properties of Twitter data along with the social relationships graph. This system goes one step further and tries to analyze user behavior across different blogging communities instead of studying only local behavior in the microblogging community. This helps to learn how and why people use such tools can be helpful in improving them and adding new features that would retain more users. In this work, they identified different types of user intentions and studied community structures. An important difference here is the frequency of updates. On average, a blogger may update his/her blog once every few days; on the other hand, a microblogger may post several updates in a single day. This study analyzed a large social network of microblogging services. Such networks were found to have a high degree of correlation and reciprocity, indicating close mutual acquaintances among users. While determining an individual user's intention in using such applications is challenging, by analyzing the aggregate behavior across communities of users, this study can describe the community intention.

Kwak et al. [13] tackle the nature of Twitter from a different angle. They actually tried to identify if Twitter contents make it valid to call it still a social network or it turns into a news propagators of a large number of editors and audiences. In other words, this paper examines the characteristic of Twitter as a strong social and news media. In order to identify influentials on Twitter, it has ranked

users by the number of followers and by PageRank and found two rankings to be similar. On the other hand, ranking by retweets differs from the previous two rankings mechanism, indicating a gap in influence inferred from the number of followers and that from the popularity of individual tweets. Further, it classified the trending topics based on the active period and the tweets and show that the majority (over 85%) of topics are headline news or persistent news in nature.

5. LIMITATIONS AND DISCUSSION

Infrastructure. One of the obvious limitations of TI [4] is ignoring all the tweets of any other query than the previously posted to the system. That means that each query with new keywords will not find any tweets to make it to its answer, and hence it should be posted again after a while of time to have a sufficiently good answer. This basically needs to know the expected query load beforehand to allow the system to satisfy its users. In addition, the system, even with this aggressive load shedding, encounters high latency for small data sizes. For example, it shows 90 ms query latency on 80K tweets. According to the latest Twitter average rate as of April 2012, 80K tweets represented data of only 20 seconds. Thus TI looks not to scale for either reasonable large time ranges or a large variety of queries.

News Extraction. One of the main problems in TwitterStand [31] is the lack of efficient techniques that are able to automatically identify news seeders, i.e., users who primarily post news stories in the blogosphere. As TwitterStand uses a pre-defined set of seeders, this needs periodic manual maintenance for the seeder list. In addition, the most prompt news reaction, which is the main motivation for the whole Twitter news extraction idea, may not be one of those known and pre-defined users. This loses the richness of crowd-wisdom and sticks the users to have updates from certain sources by providing such nice filters and interface.

Modeling. In analyzing the churn phenomena [17], the authors do not provide a mechanism to handle the lack of sufficient statistics for newly posted queries. As the nature of Twitter data content to be volatile based on the ongoing events and users interest, and that's why the churn analysis is important at the beginning, then this volatility makes the stream behavior changes all the time and needs smart techniques to cope up with this evolving. In that sense, modeling such behavior need to take into account the transition points where the trend changes from one topic to another.

In analyzing the topical properties of Twitter data [29], the authors do not use temporal properties of data to be incorporated in topical modeling. However, microblogging data is known to be temporal by nature. Thus, topical properties, as well as all other properties, changes over time, and this is important to be highlighted and taken into consideration. This will allow queries like "How does each person usage of language evolve over time?" or "How much does the distribution of substance, status, social, and style change across parts of the social network?".

In studying the differences in structure, behavior, usage of different blogging communities [11] (including microblogging), all the presented work is not automated, and hence it is a kind of theoretical study. This is definitely helpful. However, with the fast-evolving of users' usage behaviors and trends, automated tools would be needed to cope up with the fast changes.

Acknowledgments

This research was financially supported by Umm Al-Qura University. We gratefully acknowledge the support and the generosity of the University without which this study could not have been completed.

References

[1] A. Bermingham and A. F. Smeaton. Classifying Sentiment in
Microblogs: Is Brevity an Advantage? In *Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM*, pages 1833–1836, Toronto, Canada, Oct. 2010.

[2] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill, and J. Lin. Earlybird: Real-Time Search at Twitter. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, pages 1360–1369, Washington, D.C., USA, Apr. 2012.

[3] C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to Ask? Jury Selection for Decision Making Tasks on Micro-blog Services. *Proceedings of the International Conference on Very Large Data Bases, VLDB*, 5(11):1495–1506, 2012.

[4] C. Chen, F. Li, B. C. Ooi, and S. Wu. TI: An Efficient Indexing Mechanism for Real-Time Search on Tweets. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pages 649–660, Athens, Greece, June 2011.

[5] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the International Conference on World Wide Web, WWW*, pages 331–340, Raleigh, North Carolina, USA, Apr. 2010.

[6] Facebook. http://www.facebook.com/.

[7] Facebook Statistics.
http://newsroom.fb.com/Key-Facts/Statistics-8b.aspx, 2012.

[8] Foursquare. http://www.foursquare.com/.

[9] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the ACM Conference on Recommender Systems, RecSys*, pages 199–206, Barcelona, Spain, Sept. 2010.

[10] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering Geographical Topics In The Twitter Stream. In *Proceedings of the International Conference on World Wide Web, WWW*, pages 769–778, Lyon, France, Apr. 2012.

[11] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, San Jose, California, USA, Aug. 2007.

[12] S. J. Kazemitabar, U. Demiryurek, M. H. Ali, A. Akdogan, and C. Shahabi. Geospatial Stream Query Processing using Microsoft SQL Server StreamInsight. *Proceedings of the International Conference on Very Large Data Bases, VLDB*, 3(2):1537–1540, 2010.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web, WWW*, pages 591–600, Raleigh, North Carolina, USA, Apr. 2010.

[14] G. Lee, J. Lin, C. Liu, A. Lorek, and D. V. Ryaboy. The Unified Logging Infrastructure for Data Analytics at Twitter. *Proceedings of the International Conference on Very Large Data Bases, VLDB*, 5(12):1771–1780, 2012.

[15] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, pages 1273–1276, Washington D.C., Apr. 2012.

[16] J. Lin and A. Kolcz. Large-scale machine learning at twitter. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pages 793–804, Scottsdale, AZ, May 2012.

[17] J. Lin and G. Mishne. A Study of "Churn" in Tweets and Real-Time Search Queries. In *Proceedings of AAAI International Conference on Weblogs and Social Media, ICWSM*, Dublin, Ireland, June 2012.

[18] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering Spatio-temporal Causal Interactions in Traffic Data Streams. In *Proceedings of the ACM International Conference on Knowledge and Data Discovery, KDD*, pages 1010–1018, San Diego, CA, Aug. 2011.

[19] 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 *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pages 1259–1262, Athens, Greece, June 2011.

[20] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating andVisualizing Microblogs for Event Exploration. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI*, pages 227–236, Vancouver, BC, Canada, May 2011.

[21] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Processing and Visualizing theData in Tweets. *SIGMOD Record*, 40(4):21–27, 2012.

[22] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of the ACM International Conference on Management of Data, SIGMOD*, pages 1155–1157, Indianapolis, Indiana, USA, June 2010.

[23] D. McCullough, J. Lin, C. Macdonald, I. Ounis, and R. M. C. McCreadie. Evaluating Real-Time Search over Tweets. In *Proceedings of AAAI International Conference on Weblogs and Social Media, ICWSM*, Dublin, Ireland, June 2012.

[24] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM*, pages 563–572, Seattle, Washington, USA, Feb. 2012.

[25] E. Meskovic, Z. Galic, and M. Baranovic. Managing Moving Objects in Spatio-temporal Data Streams. In *Proceedings of the International Conference on Mobile Data Management, MDM*, pages 15–18, Luleå, Sweden, June 2011.

[26] G. Mishne and J. Lin. Twanchor Text: A Preliminary Study of the Value of Tweets as Anchor Text. In *Proceedings of the Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR*, pages 1159–1160, Portland, OR, June 2012.

[27] M. F. Mokbel and W. G. Aref. SOLE: Scalable On-Line Execution of Continuous Queries on Spatio-temporal Data Streams. *The International Journal on Very Large Data Bases, VLDB Journal*, 17(5):971–995, 2008.

[28] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the ACM Conference on Recommender Systems, RecSys*, pages 385–388, New York Ciy, New York, USA, Oct. 2009.

[29] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing Microblogs with Topic Models. In *Proceedings of AAAI International Conference on Weblogs and Social Media, ICWSM*, Washington, DC, USA, May 2010.

[30] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web, WWW*, pages 851–860, Raleigh, North Carolina, USA, Apr. 2010.

[31] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM GIS*, pages 42–51, Seattle, Washington, USA, Nov. 2009.

[32] V. K. Singh, M. Gao, and R. Jain. Situation Detection and Control using Spatio-temporal Analysis of Microblogs. In *Proceedings of the International Conference on World Wide Web, WWW*, pages 1181–1182, Raleigh, North Carolina, USA, Apr. 2010.

[33] Twitter. http://www.twitter.com/.

[34] Twitter Statistics. http://business.twitter.com/en/basics/what-is-twitter/, 2012.

[35] I. Uysal and W. B. Croft. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In *Proceedings of the*

ACM International Conference on Information and Knowledge Management, CIKM, pages 2261–2264, Glasgow, United Kingdom, Oct. 2011.

[36] 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 *Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM*, pages 2541–2544, Glasgow, United Kingdom, Oct. 2011.

[37] J. Yao, B. Cui, Z. Xue, and Q. Liu. Provenance-based Indexing Support in Micro-blog Platforms. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, pages 558–569, Washington D.C., Apr. 2012.

[38] D. Zhang, D. Gunopulos, V. J. Tsotras, and B. Seeger. Temporal and Spatio-temporal Aggregations over Data Streams using Multiple Time Granularities. *Information Systems*, 28(1-2):61–84, 2003.