Monitoring User Evolution in Twitter

Claudia Lauschke
Institute for Informatics
Ludwig-Maximilians University of Munich
Germany
Email: lauschkec@cip.ifi.lmu.de

Eirini Ntoutsi
Institute for Informatics
Ludwig-Maximilians University of Munich
Germany
Email: ntoutsi@dbs.ifi.lmu.de

***Abstract*—Nowadays social media are widely used for the broadcasting of different types of information, such as events, activities and opinions. Analyzing this vast amount of data for extracting models that describe individual users or groups of users has gained a lot of attention lately. In this work we analyze individual users and monitor changes in their published content over time. We propose a topic-based user profiling and monitoring approach for change detection and monitoring of profile evolution. Our method is capable of detecting persistent topics representing long term interests of the user as well as short term topics that refer to everyday events or reactions to the news. We evaluate our approach on real data from Twitter.**

I. Introduction

Blogs and microblogs like Twitter have developed to an important medium to reach the whole world. The users discuss politics, international incidents, shopping recommendations, private interests like fashion, music, movies, etc. or simply their daily lives. Nevertheless there are different types of users who use this microblogging site in different ways. Some focus on specific topics and some tweet about anything which comes in their mind. Accordingly it is also an interesting instrument for persons with strong public presence and also these users differ from each other.

In this paper, we monitor user profiles in Twitter and investigate how the profiles evolve over time. We study user evolution over time and ask questions like:

- How stable is the profile of a user over time?
- Are there any persistent topics in the profile expressing long-term interests of the user?
- Are there any random topics reflecting short term interests of the user or media influences?

The paper is organized as follows: Related work is presented in Section II. In Section III we present the user modeling at discrete timepoints. Change detection between consecutive timepoints is presented in Section IV. Our method is presented in Section V. Experiments on real data from Twitter are presented in Section VI. Section VII concluded this work.

II. Related work

Ipeirotis et al [1] study modeling and managing changes in a text database. They define as "content summary" for a database a set of keywords, weighted on their importance within the database. Meta-search services use such summaries to select appropriate databases for answering keyword-based queries. The quality of such a summary deteriorates as the contents of the database change over time. The authors propose methods to quantify and detect summary changes.

Mei and Zhai [2] study the discovery and summarization of evolutionary patterns in a text stream. They apply soft clustering with mixture models at each time period to discover latent themes and construct an evolution graph of themes by modeling theme transitions between consecutive time periods. The graph structure is used to analyze the life cycle of themes.

The MONIC framework [3] presents cluster transition modeling and detection methods and can be applied under both data and feature space changes. MONIC covers changes within a single cluster like shrink, shift etc (internal transitions) as well as changes that involve more than one cluster (external transitions), such as split and absorption, allowing insights in the whole clustering. The transition tracking mechanism of MONIC is based on the contents of the underlying data stream.

Abel et al [4] do research on user modeling strategies on Twitter. They develop a framework for describing several strategies to enrich the semantics of Twitter messages, capture personal user interests over time and relate these interests with global trending topics.

Shahaf and Guestrin [5] and Zhai, Velivelli and Yu [6] both discuss the issue of connection between articles and how to build a chain of news. In [5] the goal is to detect a chain of topics and provide an efficient algorithm to link two fixed time points. [6] propose a "Comparative Text Mining (CTM)" method which regards on common themes in a dataset of comparable articles. To find the similarities and differences of these text collections, they perform cross-collection and within-collection clustering.

Blei and Lafferty [7] study the time evolution of topics in a sequentially organized corpus of documents. By using state space models on the natural parameters of the multinomial distributions they develop a dynamic topic model. Instead of a dynamic number of topics a fixed number of topics K is set in the beginning and applied to each time slice. For each time slice for each document each word is assigned to one of these fixed topics.

Kleinberg [8] develops another approach to extract topics in document streams. As topics appearing, growing in intensity and fading away, he says that topics are signaled by 'burst of activities'. He is analyzing all words of a document stream and the sequence of their positive inter-arrival gaps. Depending on this the words are assigned to different states of intensity

for each time point. Based on a hierarchical structure of the intensity of the words, 'bursts' of words can be detected. These 'bursts' indicate topics. In contrast to our approach this one analyzes only appearing and disppearing topics and does not study the evolution of surviving topics for a longer time.

III. User Modeling

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ be a collection of documents published by a user u in a social network like Twitter. The documents constitute a stream which we monitor at discrete timepoints $\{T_1, T_2, \ldots, T_t\}$ in order to detect user content evolution over time. Let $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t\}$ be the corresponding collections of documents in each of the monitoring timepoints. E.g., $\mathcal{D}_1$ is the collection of documents published by u in the time period $[T_1, T_2)$. For each observation period T_i, we summarize the published content $\mathcal{D}_i$ and we use these summaries in order to monitor user content changes over time. A naive way for user modeling is to model the keyword distribution at each time period T_i. Such an approach though, provides only a coarse summary of the corresponding collection $\mathcal{D}_i$ and ignores any local structure in it. However, users are typically interested in a limited number of topics and publish topic related information. To this end, we propose a more elaborate modeling of the user profile at T_i in terms of the topics that might exist in the corresponding collection $\mathcal{D}_i$.

Definition 1 (User Profile): Let $\mathcal{D}_i$ be the collection of documents published during the observation period T_i by user u. Let $\{th_1, th_2, \ldots, th_j\}$ be a set of topics extracted from $\mathcal{D}_i$, such that: $th_j \subseteq \mathcal{D}_i$ and $th_1 \cup th_2 \cup \ldots \cup th_j = \mathcal{D}_i$. Intuitively, a topic is a set of documents referring to the same subject. The user profile of u at T_i, denoted by Θ_i, consists of a set of topic summaries: $(\theta_1, \theta_2, \ldots, \theta_j)$, where θ_j is the summary of the j-th topic, described in terms of:
- $(w, f(w, th_j)$ the keyword count distribution in th_j, i.e., for each keyword w, the number of documents in th_j that contain w.
- $n_j = |th_j|$, i.e., the number of documents in topic th_j.

IV. Change detection

We turn now our attention on evaluating how the content of a user changes over time. Let Θ_i, Θ_j be the topic-based profiles of user u for time periods T_i, T_j, $j > i$ respectively, extracted according to Definition 1 from the corresponding collections $\mathcal{D}_i, D_j$. The comparison between Θ_i, Θ_j relies on the comparison of their underlying topic-summaries. Let $\Theta_i = \{\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik}\}$ be the topic summaries in Θ_i and let $\Theta_j = \{\theta_{j1}, \theta_{j2}, \ldots, \theta_{jl}\}$ be the topic summaries in Θ_j. Since there might be more than one topics in Θ_i, Θ_j a kind of *mapping* or *transition* is required to define which topic of Θ_i is mapped to which topic of Θ_j. To define such a mapping a notion of distance between topics is required.

The *distance between two topics* $\theta_1 \in \Theta_i, \theta_2 \in \Theta_j$ is based on the KL-divergence of their corresponding keyword distributions (c.f., Definition 1). Formally:

$$dist(\theta_2, \theta_1) = \sum_w (\theta_2(w) - \theta_1(w)) log \frac{\theta_2(w)}{theta_1(w)} \quad (1)$$

where $\theta_1(w), \theta_2(w)$ is the probability of term w in θ_1, θ_2 respectively. This probability is given by the percentage of documents in the corresponding topic containing term w.

In order to avoid infinite distances, the two distributions should be composed of the same terms. Usually, though the distributions are not identical. To deal with such cases and also to allow for the whole vocabulary of the topics to be considered (not only the common words), smoothing is usually applied [9]. The smoothed probability of a term w in θ_2, $\theta_2'(w)$, is given by:

$$\theta_2'(w) = \begin{cases} \gamma\theta_2(w) & ,if \quad w \quad occurs \quad in \quad \theta_1 \cap \theta_2 \\ \epsilon & ,if \quad w \quad occurs \quad in \quad \theta_1 \setminus \theta_2 \end{cases}$$

where γ is a normalization coefficient estimated by:

$$\gamma = 1 - \sum_{w \in \theta_1 \setminus \theta_2} \epsilon$$

respecting the condition:

$$\sum_{w \in \theta_2} \gamma\theta_2(w) + \sum_{w \in \theta_1 \setminus \theta_2} \epsilon = 1$$

The KL scores are from 0 to infinity with 0 indicating that the two distributions are identical.

We define now the notion of evolution between two topics.

Definition 2 (Evolutionary Transition): Let θ_1, θ_2 be two topic summaries discovered during two consecutive observation periods T_i,T_j, $j > i$ respectively. There is an evolutionary transition between θ_1, θ_2 if their distance is below a given threshold δ. We denote the transition by the symbol $\mapsto$, so:

$$\theta_1 \mapsto \theta_2 : dist(\theta_1, \theta_2) \leq \delta \quad (2)$$

where $dist(\theta_1, \theta_2)$ is the topic distance function (c.f., Equation 1). In such a case, we say that θ_1 survived into θ_2 or θ_2 is evolved from θ_1.

The evolutionary transition invokes two consecutive time periods. Considering the whole observation period and the different number of topic summaries at each period, we can now define the evolution graph of the user u.

Definition 3 (Evolutionary Transition Graph): The evolutionary transition graph of a user u for a monitoring period $T_1, T_2, \ldots, T_n$ is a weighted directed graph $G = (V, E)$ where each node $v \in V$ corresponds to a topic summary θ discovered at some observation period T_i and each edge $e \in E$ indicates the existence of a survival between some topic summary θ_1 discovered in the observation period T_i and some topic θ_2 discovered in a latter observation period T_j, $j > i$. The weight of an edge is the distance between the corresponding topics.

The number of evolutionary transitions/ survivals provides and indicator of how stable or volatile is the profile of a user between two consecutive timepoints. If all clusters from one timepoint evolve into clusters of the next timepoint, the profile is stable. It there is no evolutionary transition between two consecutive timepoints, the profile is highly volatile. In the general case, some clusters will evolve and others will not evolve from a given timepoint to the next, so the user profile is subject to drift and/or shift.

V. Monitoring user profile changes

Our methodology for monitoring the evolution of a user consists of the following steps:

1) Determine the observation period (V-A).
2) Summarize user contents for each period (V-B).
3) Monitor user changes between consecutive periods (V-C).
4) Report on user changes over time (V-D).

We explain each of these steps in more detail below:

A. Determining the observation periods

Typically the user stream consists of time stamped documents (d_i, t_i) where d_i is the document published at time point t_i. A straightforward approach would be to analyze user contents in predefined points in time, e.g., every week or month. However different users publish at different rates, so applying a universal policy is not appropriate. Our approach is based on the amount of the published content for each user, so as at each observation period enough content is accumulated that allows us to extract content summaries. The algorithm detects an appropriate time window. It counts the tweets month for month beginning with the first month of the observation time until the number of tweets reaches 150 at least. This time window subdivides the whole following observation data into smaller datasets for each time period. If the first month does not deliver any tweets the following month is taken as the new starting point. This can be also applied to the following months until there is a month with more than zero tweets.

B. Summarizing user contents for each observation period

For each observation period t_i the valid collection of documents D_i for user u is derived, i.e., the documents published by the user during this period, and a summary is built upon this collection. The construction of the summary Θ_i first involves the extraction of the topics from the collection D_i and then, for each of the discovered topics, the construction of its corresponding topic summary (c.f., Definition 1). We describe both steps below.

a) Topic extraction: We use clustering for topic extraction, in particular bisecting k-Means [10]. The feature selection is based on TF×IDF. The number of the clusters is decided dynamically, based on the quality of the current clustering versus the quality of the clustering after a possible further split. In every step the cluster with the smallest distance to the cluster centroid is split into two new clusters. For computing the distance to the centroid the cosine similarity is used.

b) Topic-summary construction: For each discovered topic $\theta_j \in \Theta_i$ at time period T_i, we construct a summary based on Definition 1. In particular, we model the distribution of the keywords within the topic and also the size of the topic, i.e., the number of documents in the topic.

C. Monitor user changes between consecutive time periods

We measure the change in the user profiles between consecutive time periods t_i, t_j. To this end, we use the change detector of Section IV that maps the underlying topics in the corresponding topic-based profiles Θ_i, Θ_j based on their distance. If the distance between two topics is less than δ, an evolutionary transition/ survival is detected (c.f., Definition 2).

D. Report on user changes over time

Based on the discovered evolutionary transitions/ survivals between two consecutive timepoints t_i and t_j and on the total number of topics at each timepoint, we measure the stability of the user profile between t_i and t_j in terms of the following quantities:

- #$survivals$: the number of evolutionary transitions/ survivals from t_i to t_j. This number indicates how many of the existing clusters at t_i, had a continuation at t_j.
- #$disappearances$: the number of clusters that exist in t_i but not in t_j. This number indicates how many of the old clusters are not further continued at t_j.
- #$appearances$: the number of clusters that do not exist in t_i but exist in t_j. This number indicates how many of the clusters in t_j are new, that is, they have not been derived from some old cluster at t_i.

VI. Experiments

We experimented with a real self-crawled dataset from Twitter created as follows: We monitored a predefined list of users consisting of famous people in different fields, e.g., scientists, journalists, politicians, celebrities etc. A sample of this list is displayed in Table I. For each user, except for its name, twitter account some information are given regarding its bio and statistics about his/her "tweeting" behaviour. In particular, we mention the number of months that the user was monitored, the average number of tweets per month and the observation period (as described in V-A).

A. Monitoring user evolution results

We present representative cases for different users. We provide a short introduction to each user and then we describe his/her most persistent topics and possible exceptional topics.

1) Monitoring a politician - Barack Obama: The president of the USA uses Twitter for announcements and to discuss political and other USA related topics. We monitored him from April 2011 to September 2011 (#6 months). With an average of 28 tweets per month, a time period of 3 months was chosen as the monitoring period, resulting in 2 monitoring time periods.

There are two topics surviving both time periods, one about his political decisions and speeches and another one about his election campaigns. The first topic about *Obama's policy* contains words like white, house, Obama, president, speak, nation, support. The second topic about *Obama's campaigns* is described by words like campaign, voter, grassroot, obama. An overview of the persistent topics labels is shown in Table II.

The remaining of the clusters in both time periods also inform about Obama and his politics. So, we could conclude that this is professional account "promoting" his professional activities.

TABLE I: Description of the Selective Twitter dataset

Name	*Twitter account*	Bio	#months monitored	#tweets per month	observation period
Justin Bieber	@justinbieber	singer/pop idol	$\approx$ 16 months	300	1 month
Daniel Lemire	@lemire	CS professor	$\approx$ 12 months	75	2 months
Panos Ipeirotis	@ipeirotis	CS professor	$\approx$ 16 months	118	2 months
Barack Obama	@BarackObama	politician	$\approx$ 6 months	28	3 months
Larry King	@kingsthings	journalist/ TV show	$\approx$ 15 months	40	3 months

	Policy	**Campaigns**
April 2011 - June 2011	live manufactury obama nation house white state address economy speak affect tornado	campaign donor grassroot goal president tweet paso bring
July 2011 - September 2011	ofa president obama campaign day support speak talk action	votereg obama people voter register today president challenge

TABLE II: Topics from monitoring the profile of Barack Obama (the blue-colored notations indicate terms which re-appear in the previous or in the following period)

2) Monitoring a journalist - Larry King: Larry King is a famous US journalist and anchorman of television and radio shows. For 25 years he hosts his own show "Larry King Live" on CNN. After the last edition of the show in December 2010 he still anchored a few specials on CNN about current topics. We monitored him from October 2010 to December 2011 (#15 months). With an average of 40 tweets per month a monitoring time period containing 3 months is chosen, resulting in 5 monitoring time periods.

There is one topic surviving the whole observation time about Larry King himself, his show and his social circles. Another topic survives 3 out of 5 time periods and contains tweets about his CNN specials. The first cluster about *himself* evolves over time. In the beginning it is about the finale of Larry King's late night show. It evolves into the next cluster which describes sport events of his sons and other events Larry King attended. The next period describes his comedy tour and drifts further to other events which he attended together with his wife Shawn King. The clusters also contain tweets about his friends. Important words of these clusters are: friend, show, end, era, tour, shawnieora (the twitter nickname of Shawn King), birthday, miss. The topic about the *CNN specials* demonstrates his specials about alzheimer in May 2011, Harry Potter in July 2011 and 'Dinner with the Kings' in December 2011. It is described by words like special, alzheimer, cnn, harry, potter, dinnerwiththekings. A complete description of the persistent topics labels is shown in Table III.

Other topics Larry King tweets about are sports, politics, news and other emerging topics. Some clusters survive from one time period to another but none of these topics survives for more than two time points.

3) Monitoring a pop idol - Justin Bieber: Justin Bieber is one of the most-followed users in Twitter. He is an exceptional phenomenon of Web 2.0 because he became famous with self-recorded videos in Youtube. He uses Twitter as a channel for his promotion and announcements and makes his fans feel like taking part in his life. He publishes about 300 tweets per month which is why a monitoring time period of one month is appropriate. The whole observation time from October 2010 to January 2012 delivers 16 monitoring time periods.

There is one topic surviving 9 out of 16 time periods about Justin Bieber's love to his fans and family and also about CDs of him. Another topic which survives 5 time periods is about Bieber's movie 'Never Say Never' and a further chain with 4 consecutive clusters is describing upcoming events.

The first topic about *fans and family* is evolving over time. Beginning Bieber's appreciation of his fans and family for supporting him it drifts over to his upcoming christmas album 'Under the mistletoe'. In October the tweets are about the fans again and Bieber's shows. After this the christmas album, his fans and family become subject again. This topic chain is represented by following words: happy, proud, fan, family, support, album, love, christmas.

The second survival about *his movie* remains stable. The topic spans from November 2010 until March 2011. These clusters are defined by words like movie, nsnd (shortcut for 'never say never 3D'), neversaynever, world.

The last topic about *upcoming events* changes over time. Beginning with his concert tour 'My World Tour' the next cluster is about a rehearsal. It is followed by a cluster about shows like XFactor which he attended. The last cluster in this chain is about Justin Bieber being in the studio for new recordings. Words like show, myworldtour, neversaynever, rehearsal, tonight, studio describe this surviving and evolving topic. An overview of these persistent clusters is shown in Table IV.

There are also other shorter topic chains with not more than three time periods. Beside the already mentioned topics Justin Bieber's tweets are about his songs, his albums and his book 'First Step 2 Forever: My Story'.

4) Monitoring a professor/scientist: Daniel Lemire: Daniel Lemire is a professor at the research center in Cognitive Computer Science LICEF and at the University of New Brunswick. His research interests are Collaborative Data Management, Information Filtering and Retrieval, Database Theory, e-Learning and Data Warehousing. With an average of 75 tweets per month, a monitoring period of two months is chosen which results in total 6 monitoring time spans from

	Larry King	CNN specials
October 2010 - December 2010	end era replay sad tonight show interview itll official	
January 2011 - March 2011	time show game friend mexico star day thing egypt new	
April 2011 - June 2011	night tour time sat comedy fun atlantic open city thing	special joke book suspend tonight alzheimer dodger pick cnn week
July 2011 - September 2011	twitter com lockerz shawnieora nyc chicago game time yfrog movie	special potter harry win suspend radcliff daniel cnn sign final
October 2011 - December 2011	dinner miss time birthday terrific friend johnny king wonder interview	dinnerwiththekings talk sunday air who excited tonight pet special

TABLE III: Topics from monitoring the profile of Larry King

	Fans and family	Movie	Upcoming events
November 2010		jbdpreview neversaynever jonmchu purple glasses www show morrow hour	
December 2010		movie neversaynever incredible year crazy happy feb music come miami	
January 2011		neversaynever year week happy tomorrow swagg movie love vanity	
February 2011		movie nsnd neversaynever love today fan friend world guy inspiring	
March 2011		love tonight movie world inspiring nsnd nsnreview fan ninja pretty	
April 2011			
May 2011	love swag hope guy proud neversaynever home back tonight		
June 2011	fan love movie back support ilovemyfans night neversaynever		
July 2011	day follow happy believer swag vote justin fan		
August 2011	nsnfancut love believer tweet day neversaynever swag christmas boy belieber		
September 2011	album christmas love youtube follow swag night fan song		
October 2011	show love fan youtube mexico tonight brazil justin people day		love brazil neversaynever show people justin fan myworldtour tonight dreambig
Novemberl 2011	swag album underthemistletoe belieber mistletoe love forget song		rehearsal dreambig tonight ema excited tomorrow proud
December 2011	christmas album love underthemistletoe listen video light		tonight twitter comjustinbieber show xfactor skate nomination itv grammy
January 2012	millionbelieber justin family proud fan grow dream		studio thing tonight

TABLE IV: Topics from monitoring the profile of Justin Bieber

October 2010 to September 2011. There are 2 survivals, one over 5 out of 6 time periods about research papers and blogs, the other in 4 out of 6 time periods about programming and algorithms. The topic about *papers and blogs* begins with a cluster about publications in the field of computer science in papers and blogs which dominates all 5 consecutive time periods. The following two time periods also contain tweets about 'Wikileaks'. Words like paper, google, blog, computer, science, research, wikileaks, publish, post, work and people describe this topic. The other survival about *programming* describes algorithms and the development in Java and C++. In the following time period the topic contains tweets about database programming and google and drifts over to computer technology made by human in the last years and politics in the USA become main subject. Words like human and argument might be the reason for the connection to this different cluster. The topic chain is defined by following words: google, java, c++, alternative, algorithm, code, database, human, computer. An overview of these persistent clusters is shown in Table V.

The rest of the topics are about students and university, Wikileaks and Assange, articles, video games and world politics. No other topic chain survives more than two time periods.

5) Monitoring a professor/scientist: Panos Ipeirotis: Panos Ipeirotis is a computer science professor at New York University who uses Twitter to discuss about academics and research. We monitored him from October 2010 to January 2012 (#16 months). With an average of 118 tweets per month, a monitoring time period of 2 months was chosen, resulting in 8 monitoring time periods.

There is one topic surviving the majority of the time periods (7 out of 8), an overview of their labels is shown in Table VI. It refers to his research area, *crowdsourcing*. It contains tweets about papers and paper writing, about AmazonMechanical Turk, a crowdsourcing internet marketplace,

	Papers and blog	Programming
October 2010- November 2010	computer google people publish johndcook today science blog post	time java university class develop c++ idea gate alternative algorithm memory
December 2010 - January 2011	wikileaks research blog people book post computer paper learn education review	google write database code corporate change alternative algorithm video
February 2011 - March 2011	post blog paper time science research wikileaks read egypt work people web break	year computer human write research american paper
April 2011 - May 2011	paper google people work thing science star time usa canada open year	world country human argument future shame student thing
June 2011 - July 2011	world science publish paper code computer comwatch youtube complex wikipedia open	

TABLE V: Topics from monitoring the profile of Daniel Lemire

	Crowdsourcing
December 2010 - January 2011	mturk paper crowdsourcing spam review york idea market check work
February 2011 - March 2011	paper crowdsourcing review read work workshop article publish interest invite
April 2011 - May 2011	crowdsourcing application computer henry human meetup lesson ford april hcomp
June 2011 - July 2011	post blog crowdsourcing student paper research photo book talk cheat
August 2011 - September 2011	google review data hcomp stupid crowdsourcing room conference fake
October 2011 - November 2011	work crowdsourcing post mturk businessweek reputation lack worker google industry
December 2011 - January 2012	data busy airport video computer panosinfotech review research work product

TABLE VI: Topics from monitoring the profile of Panos Ipeirotis

and workshops/conferences. The last cluster of this topic is about his business travels. This topic is defined by words like crowdsourcing, mturk, paper, review, workshop, conference, article, research, work.

Some topics survive from one period into the next but none survives more than three periods. They describe similar subjects with different characteristics like journals, submissions and crawling data. We have to stress that the topic gaps might be also due to the fact that the specific user also tweets in Greek. In our analysis, we focus on the English tweets only.

VII. Conclusions and Outlook

By analyzing the data of the Twitter users, a few conclusions can be drawn. The users in this paper, be it journalist, professor or teen star, illustrate only a few of them but it allows perceiving a picture of the overall situation. Firstly the results show that there are different types of using Twitter profiles. Ipeirotis and Lemire focus on their research topics. Justin Bieber uses his Twitter profile primarily to promote himself. Larry King talks about his own journalism and shows. Consequently, personal information occurs in various degrees. This may lead to the conclusion that Twitter cannot be seen as a completely objective source of information even if the persons do not use Twitter as a kind of diary. Nevertheless there are also some public persons who do not publish tweets themselves which results in more objective informative profile like the one of Barack Obama.

Moreover it can be said that there is a parallel between the users and their tweets. The analysis of the examples supports this statement. Furthermore there are topic chains which are stable and remain the same but there are also topics which evolve over time. Some subjects define a user and appear in almost every time period, others emerge and soon disappear again caused by current events.

Analyzing Twitter also bears challenges. Twitter users often use urban or Internet specific language and thus it is a challenge to preprocess tweets and detect topics. Our future work focuses on enhancing the preprocessing of data in Twitter by e.g., incorporating Internet specific language and on using more elaborate methods for topic detection and topic continuation discovery. We also plan to experiment with a larger dataset of users.

References

[1] P. G. Ipeirotis, A. Ntoulas, J. Cho, and L. Gravano, "Modeling and managing changes in text databases," *ACM Trans. Database Syst.*, vol. 32, no. 3, Aug. 2007. [Online]. Available: http://doi.acm.org/10.1145/1272743.1272744

[2] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *KDD*, R. Grossman, R. J. Bayardo, and K. P. Bennett, Eds. ACM, 2005, pp. 198–207.

[3] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "MONIC: modeling and monitoring cluster transitions," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, 2006, pp. 706–711.

[4] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web," in *Proceedings of ACM WebSci '11, 3rd International Conference on Web Science (WebSci)*, 2011.

[5] D. Shahaf and C. Guestrin, "Connecting the Dots Between News Articles," in *KDD*, 2010.

[6] C. Zhai, A. Velivelli, and B. Yu, "A Cross-Collection Mixture Model for Comparative Text Mining," in *KDD*, 2004.

[7] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23th International Conference on Machine Learning*, 2006.

[8] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[9] A. Barrón-Cedeño, P. Rosso, and J.-M. Benedí, "Reducing the plagiarism detection search space on the basis of the kullback-leibler distance," in *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2009.

[10] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Text Mining*, 2000, pp. 109–111.