Extracting Personal Behavioral Patterns from Geo-Referenced Tweets

Georg Fuchs **Gennady Andrienko** **Natalia Andrienko** **Piotr Jankowski**
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven
Sankt Augustin, Germany
{firstname.lastname}@iais.fraunhofer.de

San Diego State University
Geography Annex 208
San Diego, USA
pjankows@mail.sdsu.edu

Abstract

This paper presents an exploratory study of the potential of geo-referenced Twitter data for extracting knowledge about significant personal places, behaviors and potential interests of people. The study was done analysing two months' worth of tweets from residents of the greater Seattle area.

Keywords: Visual Analytics, Spatio-temporal Analysis, Trajectories, Social Media, Twitter

1 Introduction

The high popularity of microblogging services such as Twitter in conjunction with the widespread proliferation of personal mobile devices that are able to provide location information has led to the availability of ever increasing volumes of location- and time-referenced data. For the Twitter service alone, users worldwide generate in excess of 340 million tweets each day[1]. Analysis of microblogs is interesting for a number of applications, from the validation of socio-economic theories and strategic planning [14], through localized marketing, to using Twitter users as a form of highly distributed 'social sensors' of extraordinary events or disasters.

Here we describe a visual analysis approach to extraction of information about behaviors and lifestyles of people from georeferenced Twitter posts. Both the analysis goals and the approach differ from related works that focus on the detection of extraordinary events in near-real time, e.g. an earthquake [13], but also from outwardly similar works [3] in that we focus on personal rather than general behavioral patterns.

2 Related Work

Microblogs have been investigated by researchers in computer science, social science, and other disciplines dealing with data analysis. Social scientists analyzed characteristics such as structure and relationships of social networks implied by microblogging activity [10]. Twitter in particular has been used as a source for recommendation, event detection and tracking [7] as well as sentiment [12] or hashtag analysis [6].

However, the analysis of this unstructured source is quite challenging: tweets that may be of interest to an analyst are buried in a very large amount of non-related messages, and contents of individual tweets typically contain many abbreviations, slang, typing errors and surprisingly often, just plain nonsense. This high ratio of noise combined with the brevity of individual tweets makes many traditional natural language processing tasks, such as part-of-speech tagging [9], named entity recognition [11], topic detection [15] and sentiment analysis [12] much more challenging. Yet, this kind of processing is often required to detect relevant tweets and to extract higher-level, meaningful information from them.

[1] https://business.twitter.com/en/basics/what-is-twitter/

3 Spatio-Temporal Tweet Analysis

3.1 Data Collection and Preprocessing

For the analysis presented here, we selected only tweets of two-month period (August 8th to October 8th, 2011) from the greater Seattle area in Washington State, USA. Each tweet consists of a unique tweet identifier, its geographic coordinates, time of tweeting, the tweet text itself, and an (anonymized) identifier of the Twitter user. This raw data set contains 306,326 tweets of 13,752 Twitter users (corresponding roughly to the average 1% ratio of georeferenced Tweets in the full stream).

The analysis goal of the experiment was to gain knowledge about everyday lives of people; therefore, we limited the scope of the study to the people who were present in the Seattle area for at least 10 days during the two-month period and outside of the area for less than 10 days. In addition, computer-generated messages like foursquare notifications were removed based on their fixed content pattern (e.g., "I'm at <place name> (<address>) http://..." in foursquare tweets). As the final result of the gathering and prefiltering process, we obtained a set of 163,203 individual, geo-referenced tweets from 2,607 local Twitter users.

We have constructed trajectories of the resident Twitter users from the positions of the messages by arranging the positions of each user in the chronological order.

3.2 Topic Categorization

To facilitate a semantic interpretation of spatial and temporal tweeting patterns, we enriched each Tweet with a topic categorization vector. Instead of using an automatic machine learning approach, which may lead to results that are difficult to interpret, we chose a more hands-on approach using a compiled list of 22 themes each represented by one or more keywords (cf. Table 1). For instance, the topic **family** is

Table 1. Topic category distribution over Tweets

Term	Frequency
food	6247
love	4074
family	3767
work	3076
education	2407
home	1954
private event	1928
music	1850
sports	1704
game	1678
friends	1410
health	1358
coffee	1136
transport	1120
fitness	1050
alcohol	981
weather	925
sweets	876
money	524
public event	345
tea	214
wellness	151

associated with the keywords family, mother, mom, mommy, father, dad, daddy, kids, children, son, sons, daughter, daughters, brother, brothers, sister, sisters, niece, nephew, relatives, uncle, aunt, husband, wife, folks.

We then used this set of keywords as a minimalistic ontology to categorize the tweets according to the presence of one or more topic categories. Querying the data base of 163,203 tweets for the keywords resulted in selecting 33,343 tweets (20% of the data base) containing one or more of the topic-related keywords (see Table 1).

In order to group tweets with cluster analysis, topic associations were encoded in a topic feature vector with 22 corresponding binary components attached to each tweet, which enabled abstracting from the tweets' unstructured textual content.

The two-dimensional histograms in Figure 2 show the distribution of the topics over hours of the day and days of the week. The shapes of the temporal distributions of some topics suggest that the topics of the messages may be related to the places where people are at the time of message posting (e.g. home, transport) and/or to the activities performed by the people (e.g., education, sports, eating). Hence, we can expect that by analyzing the spatial and temporal distributions and topics of the messages of different people, we can gain knowledge about their behaviors in terms of when and where they perform various activities. Additionally, by analyzing trajectories constructed from activity-specific locations distributed in time, we can learn how people move about a geographical space while tweeting about different activities.

4 Analysis of Significant Personal Places

To detect significant (i.e., repeatedly visited) personal places of the Seattle residents we clustered the positions of messages, separately for each person, by spatial proximity. We extracted 4,245 spatial clusters by means of density-based clustering [5] with the spatial distance threshold of 100 m and minimum 5 neighbors for a core point. We delineated personal places by building spatial buffers around the clusters. For each place, we aggregated the topical feature vectors of the messages posted in this place. Both the absolute counts and the percentages with respect to the total number of messages in each place were computed. A map view filtered to show only personal places with a high percentage of tweets on a specific topic gives a first overview of where this topic occupies the peoples' minds. For example, Figure 3 shows the spatial distributions of personal places with at least 20% of tweets referring to work and home, respectively, indicating

Figure 2. 2D histograms of temporal tweet distribution by topics. "All tweets" refers to the filtered data base of 33,343 tweets.

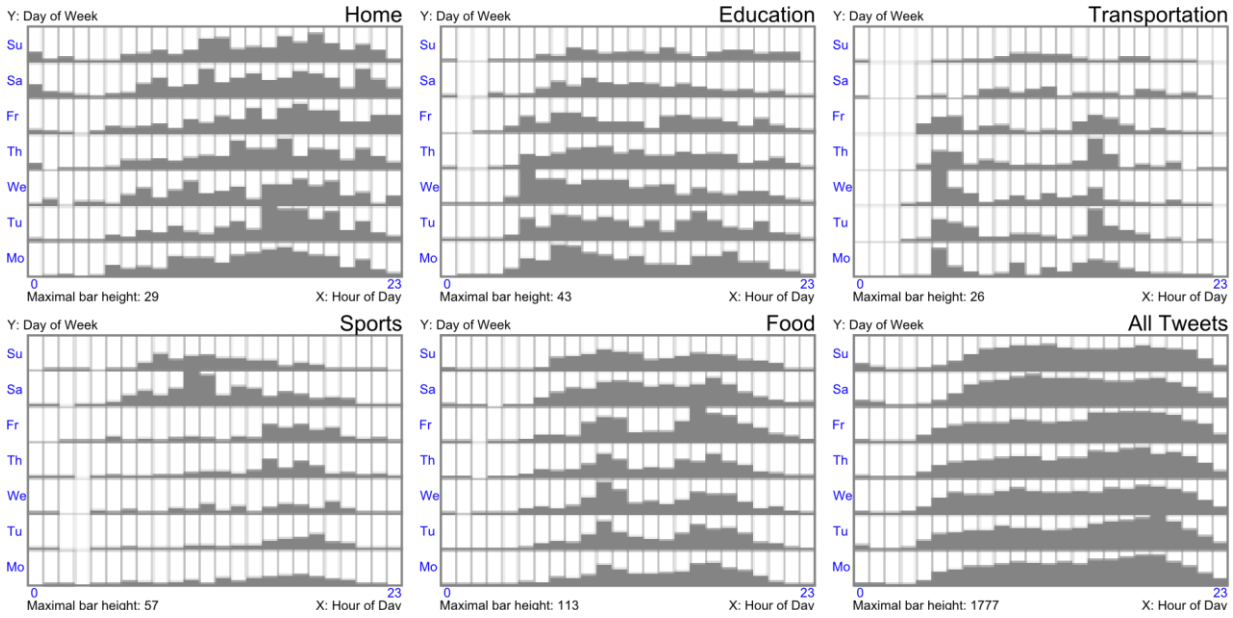
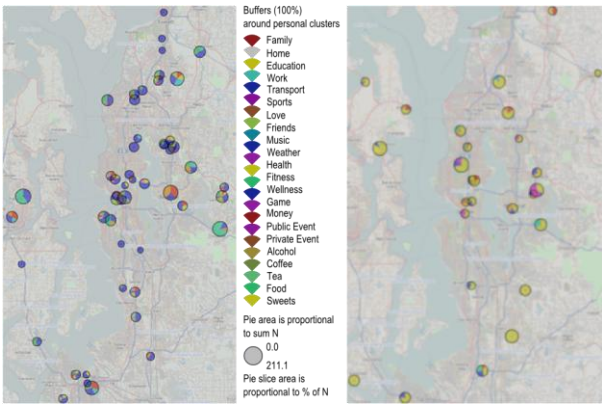


Figure 3. Spatial distribution of personal places with at least 20% of tweets referring to "work" (left) and "home" (right).



some delineation of business/industrial districts and suburban residential areas.

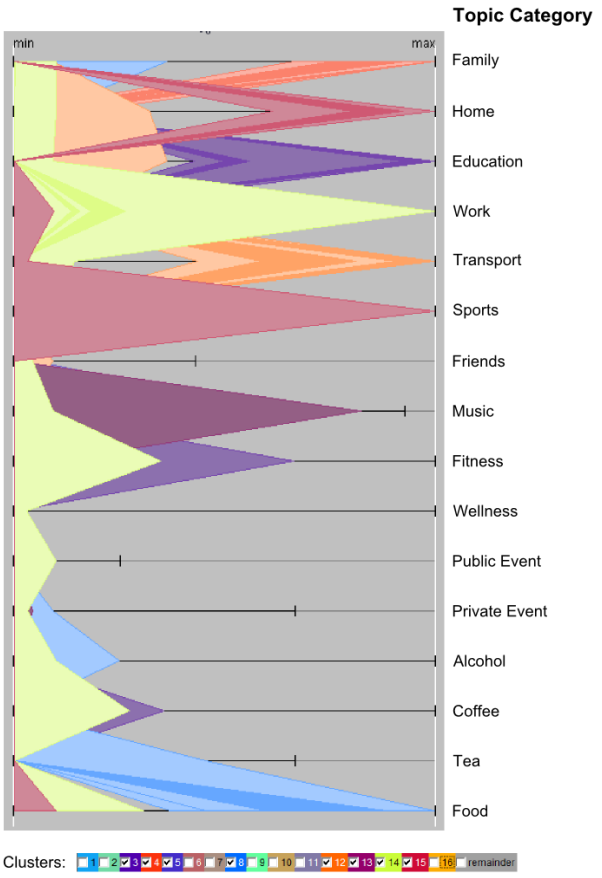
In particular, we hypothesize that the prevalence of particular topics in a personal place means that the person performs a particular type of activity in this place. The prevailing topic suggests the type of the activity.

To this end we applied k-means clustering to 2,604 places with at least 10 messages. The places are clustered according to the percentages of the different topics in the places. After several trials with different k, we chose k=16.

Some of the clusters have very clear prevalence of particular topics, as shown by the PCP in Figure 4. Here, the relative frequencies of the topics in selected clusters of places are shown in an aggregated form. Each cluster has a distinct color. Unfortunately, these clusters are quite small. Only 245 personal places out of 2,604 belong to them. The prevalent message topics in these clusters are family, home, education, work, transport, sports, music, fitness, and food. The prevailing topics suggest the possible meanings of the personal places (home, place of education, place of work, public transport stop) and/or the types of activities (studying, working, travelling, sports, entertainment, eating). There are also clusters where two or more topics are prominent and suggestive of possible activities. For example, the combination of the topics friends and family may indicate social activities.

From the temporal distribution of the person's visits to the place or tweeting, we can learn when the person performed that activity, how regularly, in what days of the week and during which times of the day. The time graphs in Figures 5 and 6 show the temporal distributions of the messages in the

Figure 4. Parallel Coordinates plot of topic prevalence (relative frequencies in % of all tweets) in clusters of places.



personal places from several selected clusters by days of the week and times of the day during working days. The time series are shown in an aggregated form by the clusters. The aggregates are colored in the same colors as in the PCP in Figure 4. The time series demonstrate that visits to the personal places and/or activities performed in these places often have specific and easily interpretable temporal patterns of occurrence. For example, sport activities as well as social activities tend to occur in the evenings of the work days and during the whole day on Saturday. Work activities occur mostly between 5:00 and 17:00 on the work days. However, there are also occurrences on the weekend, which need additional investigation. For these examples, Figure 7 shows 2D histograms of the temporal distribution of the messages, which are easier to read and interpret than the time graphs.

Figure 5. Time graph of message distribution in selected personal places (clusters) per weekday. Colors match topic categories from Figure 4.

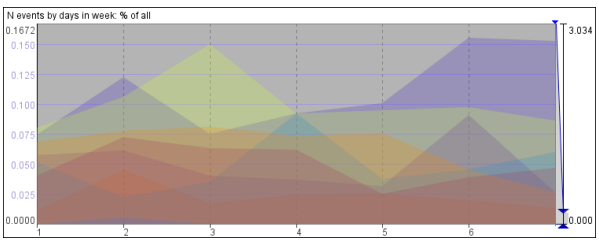


Figure 6. Time graph of message distribution in selected personal places (clusters) per hour over all working days. Colors match topic categories from Figure 4.

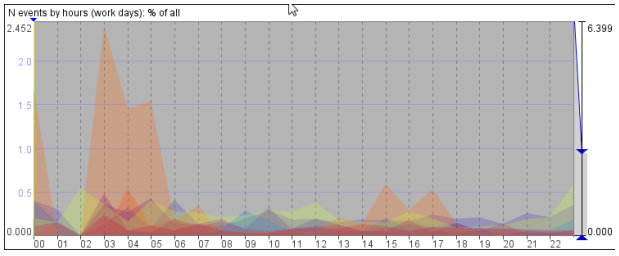
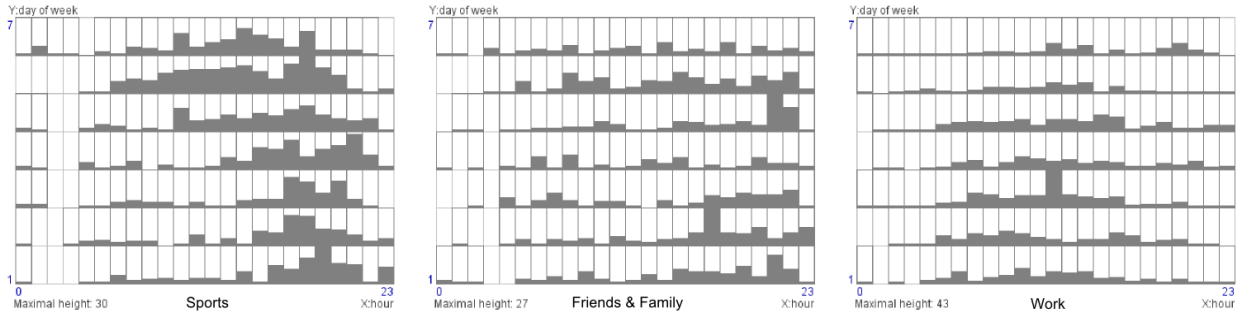


Figure 7. 2D histograms of temporal distributions of messages by hour-of-day and day-of-week for all personal places, for three selected topic categories.



Generally, personal place semantics can be established with higher certainty from the combination of message topics and temporal distributions of the messages. Thus, the places where the topic 'work' occurs only during the common work hours of the work days can be quite surely classified as work places. For the places where this topic occurs on the weekend, the distribution of this topic over the personal places of each person and over time need to be analyzed.

A spatial concentration of personal places from several people with the same attached semantic meaning serves to further increase confidence in place classification, and can also be correlated with a prevailing land use, such as office parks or residential areas.

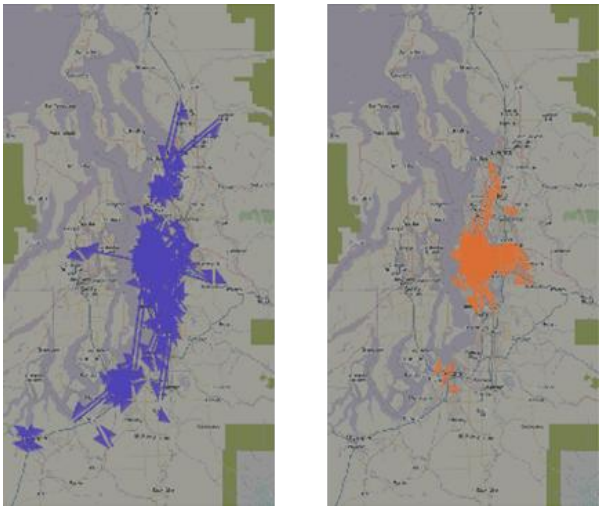
This experiment shows that for a subset of persons who regularly tweet while being in their personal places, analysis of message topics and temporal distributions allows us to attach semantic interpretation to these places and infer generic personal behavioral patterns (lifestyles) in terms of when and where persons usually perform various activities. Moreover, we can reconstruct specific personal diaries for different days.

5 Analysis of Personal Interests

For each person, we computed the absolute and relative frequencies of the different topics in the whole set of person's messages to obtain a topic fingerprint for that person. We then applied k-means clustering to the relative frequency of fingerprints and in this way discovered clusters (communities) of people with similar interests, represented by combinations of frequently occurring topics. Two examples of such combinations are illustrated in Figure 8. In particular, since several topic categories may involve similar concepts, their co-occurrence within clusters may give further insight to the nature of the common interests reflected by the clustering.

Figure 8 left shows the summarization [2] of the cluster with a moderate tendency to "work" but also a strong tendency to "food" in 60% of cases. The corresponding trajectories are widely distributed over the metropolitan region, which could be interpreted as people tweeting during or in preparation for their lunch break. Compare this to Figure 8 right, which shows the cluster of trajectories with a moderate tendency to "food" but at the same time a strong tendency to "coffee" in 40% of the cases. Here we see the trajectories concentrated in the downtown area. Quite possibly, this indicates people seeking out coffee during breaks, a habit more common to white collar workers typically found in office parks. Note that we define "moderate tendency" as the relative frequency values in the cluster being around the 3rd quartile of the value range for the whole set of trajectories, whereas "strong tendency" means relative frequencies are higher than the 3rd quartile of the value range.

Figure 8. Clustering of persons' trajectories with at least 10%-dominating topics. Left: "work" with correlated tendency to "food"; right: "food" correlating with "coffee".



For the detected communities, we can further analyze the movement behaviors, in particular, how far they can move, how many different places visit, etc. by looking at the aggregated statistics of the corresponding trajectory properties. It is also possible to find out whether people with similar interests tend to visit the same places and, if so, if the places are visited by several people simultaneously.

6 Summary

In this paper we described an experiment applying spatio-temporal aggregation and clustering methods with visual inspection to georeferenced Twitter messages to gain an understanding of significant personal places; as well as to find communities of people with similar interests and analyze their movement patterns. Tweets were first assigned topic categories using a simple taxonomy. By looking at aggregated spatio-temporal distributions of tweets from individual persons we were able to classify message location clusters according to prevailing activities performed in these significant personal places. Further, by integrating each person's tweets into trajectories and then clustering these

according to relative topic distributions, we obtained trajectory cluster representing potential communities with similar interests. Visual inspection of the spatial footprints of clusters with specific topic co-occurrences allows us to formulate hypotheses on even more specific interests.

This kind of semantic reasoning about movement patterns goes beyond spatial aggregation of movement trajectories and the detection of co-located presence of individuals, however this should be considered only a starting point for further analysis. It has been argued [8] that there exist different communities of people who co-exist in space but have different activity profiles in the temporal domain. This means that they live together but do not interact. It would therefore be interesting to examine to what extent these findings can be validated based on our data set.

Several avenues for future work emerge from this research. A long-term goal is the reconstruction of peoples' lifestyles, which has its uses in application areas such as strategic city development [14], social policy, and business analytics. A major aspect in this regard, however, is preservation of personal privacy [1].

Another aspect is the combination of visual and computational topic extraction methods. Due to their short length and often sloppy use of spelling and grammar, tweets are notoriously difficult to analyze [12]. Combining established topic modeling approaches with iterative, visual feedback-driven topic ontology might support analysis of such data.

Further experiments should extent the study to longer periods and different territories to allow comparison of patterns in space, time, as well as across population subsets and cultures; including similarity analysis between cultures and change detection with respect to a given population group over time.

Finally, there is a potential to adapt our methods to streaming settings, which would allow working directly on the Twitter stream.

References

[1] G. Andrienko and N. Andrienko. Privacy Issues in Geospatial Visual Analytics. In *Proceedings of the 8th Symposium on Location-Based Services (LBS)*, pp.239-246, 2011

[2] N. Andrienko, G. Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2011, vol. 17(2), pp.205-219

[3] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, D. Thom. Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. *Journal Computing in Science & Engineering (CiSE)*, 15(?), 2013 (accepted)

[4] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, S. Wrobel. Scalable Analysis of Movement Data for Extracting and Exploring Significant Places. *IEEE Transactions on Visualization and Computer Graphics*, 19(?), 2013 (accepted)

[5] M. Ankerst, M. Breunig, H. Kriegel, J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the ACM International Conference on Management of Data* (SIGMOD'99), pp. 49–60, 1999

[6] S. Carter, M. Tsagkias, and W. Weerkamp. Twitter hashtags: Joint Translation and Clustering. In *Proceedings of the ACM WebSci'11*, pp. 1-3, 2011

[7] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. Ebert, and T. Ertl. Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012

[8] W. Dong, B. Lepri, A. Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, ACM, 2011

[9] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of HLT'11*, pp. 42-47, 2011

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of WebKDD/SNA-KDD'07*, pp. 56-65, 2007

[11] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of HLT'11*, pp. 359-367, 2011

[12] H. Saif, Y. He, and H. Alani. Alleviating Data Sparsity for Twitter Sentiment Analysis. In *Proceedings of Making Sense of Microposts (MSM2012)*, 2012

[13] D. Thom, H. Bosch, S. Koch, M. Wörner and T. Ertl. Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. In *Proceedings of IEEE Pacific Visualization Symposium*, 2012

[14] S. Wakamiya, R. Lee, and K. Sumiya. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proceedings of the 3rd ACM SIGSPATIAL Workshop on Location-Based Social Networks* (LBSN '11), pp. 77–84, 2011

[15] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval* (ECIR'11), pp. 338–349, 2011