On the Correlation Between Textual Content and Geospatial Locations in Microblogs

Irena Grabovitch-Zuyev
Technion – Israel Institute of Technology, Haifa, Israel
siraz@cs.technion.ac.il

Yaron Kanza
Jacobs Technion-Cornell Innovation Institute
New York, USA
kanza@cornell.edu

Elad Kravi
Technion – Israel Institute of Technology, Haifa, Israel
ekravi@cs.technion.ac.il

Barak Pat
Technion – Israel Institute of Technology, Haifa, Israel
barakpat@cs.technion.ac.il

ABSTRACT

Microblogs allow users to publish geo-tagged posts—short textual messages assigned to a geographic location. Users send posts from places they visit and discuss an idiosyncratic mixture of personal and general topics. Thus, it is reasonable to assume that the locations and the textual content of posts will be unique and will identify the posting user, to some extent. This raises the question whether there is a correlation between the locations of posts and their content. Are users who are similar from the geospatial perspective (i.e., who send messages from nearby locations) also similar from the textual perspective (i.e., send messages with similar textual content)? Do posts with similar content have a spatial distribution similar to that of any random set of posts? We present a study that focuses on these questions. We provide statistical tests to examine the correlation between textual content and geospatial locations in tweets. We show that although there is some correlation between locations and textual content, they provide different similarity measures, and combining these two properties for identification of users by their posts outperforms methods that merely use locations or only use the textual content, for identification.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Experimentation, Measurement

Keywords

Microblogs; geospatial/textual similarity; correlation

GeoRich'14 June 23 2014, Snowbird, Utah, USA

http://dx.doi.org/10.1145/2619112.2619115.

1. INTRODUCTION

Microblogs, such as Twitter, are important and agile tools for expeditiously sharing real-time information among people. The information submitted on Twitter is typically personal in its nature, however, large sets of tweets frequently reflect popular trends, events of different types, and various phenomena. For example, the popularity of the hashtag `#ladygaga` reflects a musical trend; "Hurricane Sandy" is an event with many mentions in tweets during its occurrence; and having more tweets sent from Manhattan, New York than from Greenland is a general phenomenon.

Discovering trends, events or phenomena by analyzing microblog posts is an area that receives a growing attention, due to the easy access to the data and the fact that the data are received from many independent sources—having multiple independent sources typically increases the reliability of the information and obstructs manipulations. Another main advantage of microblogs is that data are up-to-date and reflect current developments or real-time events [16,19].

Typically, microblogs allow users to post geo-tagged messages from different locations. Each geo-tagged post contains both textual content and the location from which it was sent. Since users typically post messages from places they frequently visit, the set of locations of the posts of a user is almost always unique. Also, the contents of posts are unique. In this paper we show this by selecting an arbitrary user, partitioning the posts of the user, randomly, into two parts and comparing the distance between the parts to the distance between one of these parts to sets of posts of other users. To do so, we use common distance measures—nearest-neighbor distance as a distance based on locations and TF/IDF (cosine similarity) as a distance based on content. Our experiments show that both locations of posts and the textual content of posts identify people.

Knowing that both locations and content of posts identify users raises the question whether there is a correlation between these two attributes. Does a location proximity between posts of users indicate a similarity between the contents of their posts? Does similarity between contents of posts result in location proximity of the posts? These questions have practical implications. Detection location similarity is relatively easy, so if there is a correlation between the locations and content, it will be efficient to find users

whose posts have near locations when trying to discover users whose posts have similar content. Also, if there is a correlation between content and locations, we can learn about the whereabouts of a user whose posts are not geo-tagged but with content similar to the content of posts of a user who does geo-tag her posts. On the other hand, if there is no correlation between locations and content of posts, these two aspects of posts could be combined to improve the discovery of similarity among users.

To examine if there is a correlation between locations and content of posts, we conducted statistical tests to compare two rankings of users, according to their similarity to a selected user—in one ranking we used similarity based on locations, and in the other ranking, similarity based on content. We used two statistical tests to examine the correlation: *(1)* a test that compares our results to the hypergeometric distribution of the intersection of independent subsets, and *(2)* the *Spearman's rank correlation coefficient*, which examines how far is the ranking of one variable from being a monotone function of the other variable. We then propose two methods to combine locations with content and we show the improvement gained.

The correlation tests described above are with respect to arbitrary users. To complete the examination, we also applied tests that do not rely on the association of posts to users. In one group of experiments, we choose frequent terms from posts and examined whether posts containing these terms are spatially distributed as a randomly selected set of posts. We did so by measuring the *spatial variance* of the posts with respect to the center-of-mass of the locations. In complimentary experiments, we tested if for small areas, the posts sent from these areas contain terms that are somewhat unique for the area. We show that there are topics that are related to a specific location (e.g., "Rockefeller Center") and topics that are unrelated to any specific location (e.g. "Selfie"). We illustrate in statistical tests the difference between these two types of terms and show that their existence affects the correlation between locations and content. Nonetheless, in all cases, the combination between locations and content can improve the identification of users.

2. RELATED WORK

The connection between social networks and locations is receiving a growing attention [5–7, 17, 18]. Recently there has been a growing interest in using microblogs for detecting earthquakes and assessing the aftermath [15, 22, 23, 32]; and for detecting the spread of epidemics and pandemics [12]. It has been shown how to use Twitter as a news source [24, 28] or for detecting events [1, 11, 20, 29]. Some studies focused on using microblogs for recommendations [14, 27], reasoning about urban activities [8, 25] and discovering communities [3, 4]. Some papers studied detection of topics in posts within a bounded geographic region [10, 13, 30], however, all these papers did not thoroughly study the correlation between locations and content in posts.

3. FRAMEWORK

We present now our framework. A *post* is a pair $p = (l, c)$ of location l and textual content c—the textual content is a list of terms. Locations are points on a sphere (Earth) and the distance between two posts, denoted *dist*, is the Haversine distance between their locations. Each post is related to the user who submitted it. Hence, we associate with each user u the set P_u of posts of u. When it is clear from the context, by referring to u we consider P_u. To measure the similarity between users, we measure the spatial and the textual distances between the posts of the users.

Spatial Similarity. The *spatial distance* between two users is the average distance between pairs of locations in the *nearest-neighbor* matching between the sets. Formally, let P and P' be two sets of posts, where $l_1, \ldots, l_n$ are the locations of the posts in P and $l'_1, \ldots, l'_{n'}$ are the post locations of P'. We denote by $|P|$ the number of posts in P. A nearest-neighbor matching $\mu : P \to P'$ is a function that maps each post of P to its nearest neighbor in P', i.e., $\mu(p_i) = p'_j$ *if* $dist(p_i, p'_j) \leq dist(p_i, p'_{j'})$ *for all* $p'_{j'} \in P'$. We denote by $dist(p_i, \mu(p_i))$ the distance between post p_i to its nearest neighbor in P'. For two sets of posts P and P' such that $|P| \geq |P'|$, the spatial distance between them is the average nearest-neighbor distance $\left(\sum_{i=1}^{n} dist(p_i, \mu(p_i))\right)/n$. If one set is larger than the other, then the mapping is from the larger set to the smaller one because our experiments show this provides better results than mapping the smaller set to the larger one. The *spatial similarity* between sets is the inverse of the spatial distance.

Content Similarity. For measuring content similarity, we use the *terms vector space model* [2]. Each user is represented by a vector of terms, defined with respect to a *corpus*. The *corpus* is the set of terms (unigrams) that appear in the content of at least one post. Phone numbers, digit sequences, one letter words and stop words (e.g. "and", "is", "of") are discarded. The terms are added to the corpus lower-cased and stemmed using the Porter Stemmer [21]. The corpus also contains the following statistics on terms: *(1)* the number of users that used the term in their posts; *(2)* the number of posts that contain the term; and *(3)* the number of appearances of the term in all the posts.

The terms vector of a user assigns a weight to each term in the corpus, to indicate the relevance of the terms to the user. The weights were calculated using TF-IDF. For a term t and user u, $TF(t)$ is the number of times t appears in the posts P_u. By $DF(t)$ we denote the ratio of the number of users who used t in at least one post to the total number of users, and $IDF(t) = \log\left(DF(t)^{-1}\right)$. The weight $w_u(t)$ of a term t is defined as $TF(t) \times IDF(t)$. This reflects the intuition that unique terms provide more information than common terms and that the frequency of terms is also significant.

The *content similarity score* of two users is measured as the cosine of the angle between the terms vectors of the users. Let N denote the number of terms in the corpus and t_i be term i. Consider two users u and v. Then, the content similarity of u and v is

$$\frac{u \cdot v}{|u||v|} = \frac{\sum_{i=1}^{N} w_u(t_i) \cdot w_v(t_i)}{\sqrt{\left(\sum_{i=1}^{N} w_u(t_i)\right)^2}\sqrt{\left(\sum_{i=1}^{N} w_v(t_i)\right)^2}}.$$

Spatial Variance. The *spatial variance* of a set P of posts is a measure of the spread of the posts in the investigated area. It is measured with respect to the *center of mass* of P. Essentially, the center of mass of P is the point c whose X coordinate is the arithmetic mean of the X coordinates of the posts of P, and whose Y coordinate is the arithmetic mean of

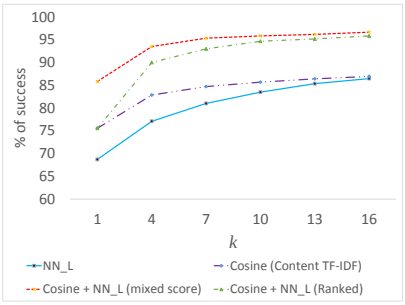

Figure 1: Accuracy as a function of the relaxation level k, in the identification test.

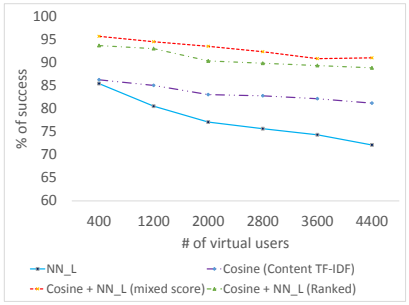

Figure 2: Accuracy as a function of the number of candidate users, in the identification test.

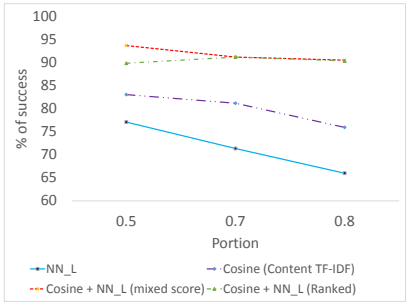

Figure 3: Accuracy as a function of the partition of the user u (portion), in the identification test.

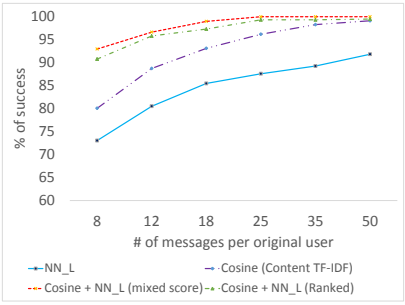

Figure 4: Accuracy as a function of the number of posts P_u of the user u, in the identification test.

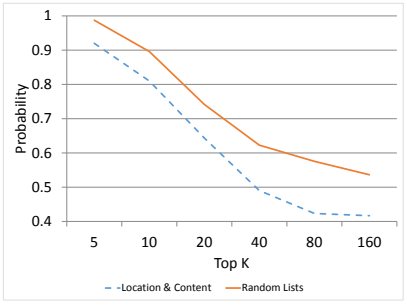

Figure 5: Hypergeometric cumulative probability of the size of the intersection of the top-k users of the ranked lists.

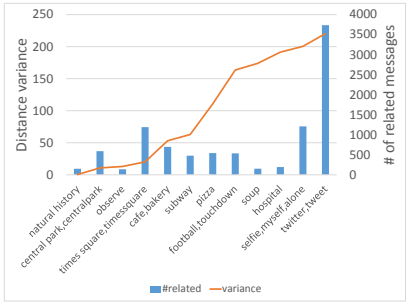

Figure 6: Distance variance (orange line) for messages on various topics. Blue bars indicate the number of messages of the topic.

the Y coordinates of the posts of P. The *spatial variance* of P is defined in the usual way, i.e., $\left(\sum_{p \in P} dist(p,c)^2\right)/|P|$.

User Identification Test. A *unique* attribute is an attribute that can be used to identify a user fairly well. *User name* is an example of a unique attribute, whereas the language used to write the posts is typically not unique. Identification of users has many applications in privacy, data integration, recommendations (e.g., recommendations can be improved by considering data about a user from two different social networks rather than from just one), etc.

To test whether some attribute is unique, we apply the following test. We take a user u and N additional users. We randomly partition the posts P_u into two sets P_{u_1} and P_{u_2}. We consider u_1 and u_2 as *virtual users* although these are two parts of u. A unique attribute can be used to detect u_2 as the other part of u_1 among $N \cup \{u_2\}$. We use spatial similarity or content similarity to find the user most similar to u_1 among the users of $N \cup \{u_2\}$, to test the ability to identify a user based on post locations or content. Note that we deliberately chose common techniques for this task because we test the attributes rather than the technique.

To test the quality of an attribute in the identification task, we choose m arbitrary users, run the identification test for each one of them, with respect to an arbitrary set of N users, and report the ratio of success. A high ratio indicates a good ability to identify the users.

Combining Attributes for Identification. An important question we examine is to what extent the combination of locations and content helps in identifying users. To test this, we consider two methods of combining the attributes.

Consider a user u. The spatial similarity and the content similarity measures, each ranks the other users according to their similarity to u. Thus, in a combination of the attributes, we have two ranked lists L_1 and L_2 of users. Let $pos_j(v)$ be the location of user v in list L_j ($j \in \{1,2\}$), e.g., if v is the first element in L_1, then $pos_1(v) = 1$, if it is the second, then $pos_1(v) = 2$, and so on. In a *rank-based combination* of the lists, we assign to each user v the values $RBC(v) = \min\{pos_1(v), pos_2(v) + \frac{1}{2}\}$. Adding $\frac{1}{2}$ to the positions of the second list guarantees that each user receives a unique score, i.e., for any two users $u_1 \neq u_2$, holds $RBC(u_1) \neq RBC(u_2)$—if the scores are from a single list, the two users have different positions, and if the scores are from different lists, then only one of them is an integer. Finally, we simply sort the users according to the RBC scores.

Another approach to combine the attributes is to rank users according to the average of the scores. To that end, the location-based similarity scores are normalized to be in the range $[0,1]$. (Content-based scores are already normalized.) Then, each user v is assigned the arithmetic mean of the spatial-similarity and content-based similarity scores.

4. CORRELATION TESTS

We present now an experimental study of the correlation between locations and content in sets of tweets, including results of identification tests and of correlation tests.

Topic	# of posts	variance
natural history	150	0.3
rockefeller	646	7.25
central park	590	10.73
observe	136	12.96
times square	1189	20.13
china town	122	28.49
cafe, bakery	698	53.07
subway	477	62.85
pizza	542	110.67
Random (150)	150	160.1
football, touchdown	534	163.50
soup	152	173.89
Random (600)	600	176.15
Random (300)	300	180.17
Random (1500)	1500	191.07
hospital	193	191.34
Random (1000)	1000	199.96
selfie, myself, alone	1208	200.43
twitter, tweet	3737	220.28

Table 1: Spatial variance of different topics. The Random (X) topics are a random selection of X posts, serving as a yardstick.

4.1 Identification Tests

Our first set of experiments executed the identification test to compare the identification effectiveness of locations, content and the combination of these two attributes. Our dataset consisted of 204,860 tweets of 22,900 users, posted from the area of Los Angeles, California. In the tests, we only used users with at least ten tweets. In each experiment, we chose N random users. Then we selected from them a user u. We partitioned each user into two parts, and in particular, partitioned the posts P_u of u into two sets P_{u_1} and P_{u_2}. Let $U_{1/2}$ denote the set of $2N$ partitioned users. The *portion* of the partition is the ratio $\rho = |P_{u_1}|/|P_u|$. The portion tells if the partitions were into equal parts, i.e., the portion is 0.5, or to unequal parts, e.g., the portion is 0.3. We relaxed the tests by choosing a value k, referred to as *relaxation level*, such that a success is when the virtual user u_2 is among the k users most similar to u_1, in the set $U_{1/2} \setminus \{u_1\}$, i.e., in the set comprising the $2N-1$ parts of the chosen N users, except for u_1 (note that this set includes u_2). For each N users, we applied the identification test for each one of the N users, and we applied this three times using different sets of N random users. Hence, the reported values are an average of $3N$ identification tests. The default values are $N = 1000$, $k = 4$, $\rho = 0.5$.

In the experiments, we compared four methods. As a location-based method we used the nearest-neighbor similarity measure, denoted NN_L. As a content-based method we used the cosine similarity method (TF-IDF). For the combination of the methods, we used the combination that mixes the scores (arithmetic mean of the scores) and the combination based on the rankings.

In Fig. 1 we present the results of an experiment that tested the accuracy of the identification as a function of k (i.e. as a function of the relaxation level). It can be seen that identification based on content is slightly better than identification based on locations, especially for small values of k. Combining the attributes improves the results where combination based on scores outperforms combination based on ranking (note that for $k = 1$, combination based on ranking merely takes the first user from the content-based ranking).

Figures 2, 3 and 4 validate the results of Fig. 1 by varying different parameters. In Fig. 2, the accuracy is measured as a function of the number N of considered users. In Fig. 3, the accuracy is measures as a function of the portion ρ. Fig. 4 presents the accuracy as a function of the number of posts in P_u. Obviously, when there are more messages in P_u or in the two partitions P_{u_1} and P_{u_2}, identification becomes easier (e.g., when $\rho = 0.8$, one of the virtual users has size $0.2|P_u|$ and thus, it is relatively small, so this case is harder than $\rho = 0.5$). Clearly, when N grows, the problem becomes more difficult. An interesting conclusion, however, is that in all these cases, identification based on content is better than identification based on locations, and combining the rankings improves the identification. Already for $|P_u| = 25$, the combination achieves nearly 100% accuracy.

Both methods of combining locations with content gained better results than the methods that only use one attribute. Combining the scores is marginally better than combination based on ranks. To assert this assumption—that combination based on scores is better than combination based on ranks—we performed a *Wilcoxon signed-rank* statistical test [31]. This test showed we can deny the null hypothesis with $\alpha = 0.00014$, so the statistical significance of the assumption is high.

4.2 Correlation Tests

To examine the correlation between ranking based on locations and ranking based on content, we conducted the following tests. We randomly selected 200 users, and for each user u among these 200 users, we chose other 1000 random users. We then ranked these 1000 users according to their similarity to u, using both location similarity and content similarity. This provided a pair of ranked lists, for each user u—one list where the users are ranked according to location similarity and another list where the users are ranked according to content similarity. Then, we applied two correlation tests on each such pair of lists. In the first test, we computed the Spearman's rank correlation coefficient of the two lists [26]. The average value of the coefficient in all these tests was practically zero. This indicates there is no dependency between the order in one list to the order in the other list. Note that this test is very strict as it measures differences in the positions of users in the two lists.

A more permissive test we conducted is based on using the hypergeometric cumulative probability of finding shared users among the k top users of the two lists. In the experiment, for different k values we computed the intersection of the top-k users of the two lists. Then, we computed the probability of achieving an intersection that is at least as large as the actual intersection, using a hypergeometric cumulative probability. When the intersection of the two lists is small, the hypergeometric cumulative probability is high, and as the size of the intersection grows, the probability decreases. The results of this test, using $N = 1000$ users, are presented in Fig. 5. The blue dashed line is the hypergeometric cumulative probability of the size of the intersection of the top-k users of the location-based ranking and the content-based ranking. The solid orange line is the probability achieved for two lists of 1000 users randomly sorted (averaged over 1000 runs). It can be seen that the curve of the intersection of the content-based and location-based

Figure 7: Selected areas in Manhattan, NYC, presented using OpenStreetMap [9].

rankings is below the curve of the intersection of two random lists. This shows that the intersection of the top-k users of the location-based ranking and the content-based ranking has a larger size than the intersection of random lists, and the probability to achieve such an intersection of the ranked lists by chance is low. So, there is a correlation between the ranks, although it is not apparent from the order of the lists. That is, when selecting k users who are the most similar to a user u based on content similarity, these users are expected to also be more similar to u from the spatial perspective than a random set of users, and vice versa.

4.3 Spatial Variance

The tests reported so far are based on comparisons of users. Next, we report the results of experiments to test the spatial variance of messages related to different topics. We used 205,000 tweets posted from Manhattan, NYC. We selected several topics—each topic is specified as a set of terms. A post is considered related to the topic if it contains one of the terms of the topic. We computed the location variance of the posts of these topics. As a yardstick, we also randomly selected sets of of posts, with various sizes, and computed their location variance. The results are presented in Table 1 and in Fig. 6. It can be seen that there are topics which are related to a location, e.g., "natural history" or "observe" while other topics are not associated with a specific location, e.g., "hospital" or "Selfie". Applying such test on the terms a user uses can provide some indication to places the user visited if the used terms have a small location variance, however, the content of posts cannot be used for detecting visited places when the location variance is big.

4.4 Unique Terms per Area

We tested whether there are locations that are characterized by specific topics. To test this, we selected arbitrary locations in Manhattan, NYC and extracted the terms that appeared in posts submitted from these areas. We report the results for the areas presented in Fig. 7. For the most frequent terms in these areas we measured the following parameters. *Frequency* indicates how many posts, among those sent from the area, contain the term. *Percentage* is the percentage of the posts containing the term, among the posts sent from the area. *Ratio* stands for the percentage of the posts in the area containing the term, divided by the percentage of the posts in all NYC containing the term. The results are presented in Table 2.

Frequent terms in Area 1					
Word	New	Time	NY	Love	Back
Frequency	12	8	8	8	6
Percentage	10.4	7	7	7	5.2
Ratio	1.69	2.61	1.48	1.89	3.14

Frequent terms in Area 2					
Word	New	NY	York	Time	Manhattan
Frequency	40	29	22	17	13
Percentage	11.9	8.6	6.6	5	3.8
Ratio	1.93	1.84	1.67	1.9	11.2

Frequent terms in Area 3					
Word	Rockefeller	Center	New	York	Tree
Frequency	548	499	206	191	162
Percentage	53.4	48.6	20	18.6	15.7
Ratio	171.88	62	3.25	4.73	93.20

Frequent terms in Area 4					
Word	Central	Grand	Terminal	Other	New
Frequency	192	189	169	106	93
Percentage	47.5	46.7	41.8	26.2	23
Ratio	111.33	217.9	156	7.9	3.73

Table 2: Frequent terms in Areas 1-4. Additional notable terms in Area 3 are "Christmas", "Observation" and "Nintendo" with ratios of 35, 143 and 137.

The tables show that an area containing Rockefeller Center (Area 3) and an area containing Grand Central Terminal (Area 4) have significant terms with high percentage and high ratio. Such terms can typify the areas. Area 1 and Area 2 do not have such characterizing terms. Thus, for some areas, many messages sent from them use unique terms whereas for others, there are no unique terms to distinguish them. Note that the characterizing terms can be used to point out *interesting places*, e.g., by seeing the term "Nintendo" in Area 3, one can learn that there is a significant place related to Nintendo in this area (indeed "Nintendo World Store" is located in the area). The terms "Christmas" and "Tree" indicate a noteworthy Christmas Tree in Area 3. Hence, in many cases, such term extraction and the computation of the ratio can be used to discover enthralling places and improve our understanding of urban districts.

5. CONCLUSIONS

In this paper we study relationships between locations and content of microblog posts. We show that when considering users, similarity based on locations is different from similarity based on content. Actually, these two attributes can be combined to improve the identification of users. However, by studying posts without considering users, we see that many terms can be associated with specific places (i.e.,

many of the posts containing these terms are sent from a small area) while other terms are not associated with a specific area. Similarly, there are areas that are characterized by the topics of their posts, while other areas do not have specific terms associated with them. This study points out that the correlation between locations and content in microblog posts is complex, however, understanding it can be useful in different domains and for various applications, such as recommendation systems, privacy and analysis of trends.

6. ACKNOWLEDGMENTS

This research was supported in part by the Israel Science Foundation (Grant 1467/13) and by the Isreali Ministry of Science and Technology (Grant 3-9617).

7. REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proc. VLDB Endow.*, 6(12):1326–1329, Aug. 2013.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[3] T. V. Canh and M. Gertz. A spatial lda model for discovering regional communities. In *Proc. of the 2013 IEEE/ACM International Conf. on Advances in Social Networks Analysis and Mining*, pages 162–168, 2013.

[4] M. De Choudhury, N. Diakopoulos, and M. Naaman. Unfolding the event landscape on twitter: Classification and exploration of user categories. In *Proc. of the ACM 2012 Conf. on Computer Supported Cooperative Work*, pages 241–244, 2012.

[5] Y. Doytsher, B. Galon, and Y. Kanza. Querying geo-social data by bridging spatial networks and social networks. In *Proc. of the 2nd ACM SIGSPATIAL Inter. Workshop on Location Based Social Networks*, pages 39–46, 2010.

[6] Y. Doytsher, B. Galon, and Y. Kanza. Storing routes in socio-spatial networks and supporting social-based route recommendation. In *Proc. of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 49–56, 2011.

[7] Y. Doytsher, B. Galon, and Y. Kanza. Querying socio-spatial networks on the world-wide web. In *Proc. of the 21st International Conf. Companion on World Wide Web*, pages 329–332, 2012.

[8] N. Gnanasambandam, K. Thompson, I. F. Ho, S. Lam, and S. W. Yoon. Towards situational pattern mining from microblogging activity. In *Proc. of the 21st International Conf. on World Wide Web*, pages 661–666, 2012.

[9] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

[10] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proc. of the 21st International Conf. on World Wide Web*, pages 769–778, 2012.

[11] E. Ilina, C. Hauff, I. Celik, F. Abel, and G.-J. Houben. Social event detection on twitter. In *Proc. of the 12th Inter. Conf. on Web Engineering*, pages 169–176, 2012.

[12] N. Kanhabua, S. Romano, A. Stewart, and W. Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proc. of the 21st ACM International Conf. on Information and Knowledge Management*, pages 2686–2688, 2012.

[13] K.-S. Kim, R. Lee, and K. Zettsu. mTrend: Discovery of topic movements on geo-microblogging messages. In *Proc. of the 19th ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*, pages 529–532, 2011.

[14] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. LARS: A location-aware recommender system. In *Proc. of the 2012 IEEE 28th International Conf. on Data Engineering*, pages 450–461, 2012.

[15] Y. Liang, J. Caverlee, and J. Mander. Text vs. images: On the viability of social media to assess earthquake damage. In *Proc. of the 22Nd International Conf. on World Wide Web Companion*, pages 1003–1006, 2013.

[16] P. Liu, J. Tang, and T. Wang. Information current in twitter: Which brings hot events to the world. In *Proc. of the 22Nd International Conf. on World Wide Web Companion*, pages 111–112, 2013.

[17] M. F. Mokbel and M. Sarwat. Mobility and social networking: A data management perspective. *Proc. VLDB Endow.*, 6(11):1196–1197, Aug. 2013.

[18] M. Naaman. Geographic information from georeferenced social media data. *SIGSPATIAL Special*, 3(2):54–61, 2011.

[19] M. Okazaki and Y. Matsuo. Semantic twitter: Analyzing tweets for real-time event notification. In *Proc. of the 2008/2009 Inter. Conf. on Social Software: Recent Trends and Developments in Social Software*, pages 63–74, 2010.

[20] A.-M. Popescu, M. Pennacchiotti, and D. Paranjpe. Extracting events and event descriptions from twitter. In *Proc. of the 20th International Conf. Companion on World Wide Web*, pages 105–106, 2011.

[21] M. F. Porter. An algorithm for suffix stripping. In K. Sparck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[22] B. Robinson, R. Power, and M. Cameron. A sensitive twitter earthquake detector. In *Proc. of the 22Nd International Conf. on World Wide Web Companion*, pages 999–1002, 2013.

[23] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proc. of the 19th International Conf. on World Wide Web*, pages 851–860, 2010.

[24] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *Proc. of the 17th ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*, pages 42–51, 2009.

[25] C. Sengstock, M. Gertz, H. Abdelhaq, and F. Flatow. Reliable spatio-temporal signal extraction and exploration from human activity records. In *Proc. of the 13th International Conf. on Advances in Spatial and Temporal Databases*, pages 484–489, 2013.

[26] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

[27] Y. Takeuchi and M. Sugimoto. Cityvoyager: An outdoor recommendation system based on user location history. In *Proc. of the 3rd International Conf. on Ubiquitous Intelligence and Computing*, pages 625–636, 2006.

[28] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: A new view on news. In *Proc. of the 16th ACM SIGSPATIAL International Conf. on Advances in Geographic Information Systems*, pages 18:1–18:10, 2008.

[29] K. N. Vavliakis, A. L. Symeonidis, and P. A. Mitkas. Event identification in web social media through named entity recognition and topic modeling. *Data Knowl. Eng.*, 88:1–24, Nov. 2013.

[30] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *Proc. of the 35th European Conf. on Advances in Information Retrieval*, pages 356–367, 2013.

[31] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83, 1945.

[32] J. Yin, S. Karimi, B. Robinson, and M. Cameron. ESA: Emergency situation awareness via microbloggers. In *Proc. of the 21st ACM International Conf. on Information and Knowledge Management*, pages 2701–2703, 2012.